



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Agronomia Animali Alimenti Risorse Naturali e Ambiente (DAFNAE)

CORSO DI DOTTORATO DI RICERCA IN SCIENZE DELLE PRODUZIONI VEGETALI

CICLO: XXX

**EXPLOITING GENOMICS AND MOLECULAR MARKERS FOR PLANT
GENETICS AND BREEDING**

Coordinatore: Ch.mo Prof. Sergio Casella

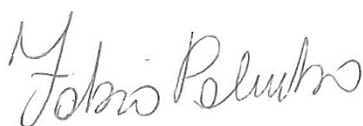
Supervisore: Ch.mo Prof. Gianni Barcaccia

Dottorando: Fabio Palumbo

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

(signature/name/date)



Fabio Palumbo 10/01/2018

A copy of the thesis will be available at <http://paduaresearch.cab.unipd.it/>

Dichiarazione

Con la presente affermo che questa tesi è frutto del mio lavoro e che, per quanto io ne sia a conoscenza, non contiene materiale precedentemente pubblicato o scritto da un'altra persona né materiale che è stato utilizzato per l'ottenimento di qualunque altro titolo o diploma dell'università o altro istituto di apprendimento, a eccezione del caso in cui ciò venga riconosciuto nel testo.



(firma/nome/data)

Fabio Palumbo 10/01/2018

Una copia della tesi sarà disponibile presso <http://paduaresearch.cab.unipd.it/>

Index

Riassunto	1
General abstract	3
Critical aspects on the use of microsatellite markers for assessing genetic identity of crop plant varieties and authenticity of their food derivatives	6
Developing a molecular identification assay of old landraces for the genetic authentication of typical agro-food products: the case study of the barley ‘Agordino’	46
Venetian local corn (<i>Zea mays</i> L.) germplasm: disclosing the genetic anatomy of old landraces suited for typical cornmeal mush production	74
The leaf transcriptome of fennel enables the characterization of the t-anethole pathway and the discovery of microsatellites and single-nucleotide variants	100
First draft genome sequencing of fennel (<i>Foeniculum Vulgare</i> Mill.): identification of simple sequence repeats and their application in marker-assisted breeding	132
Construction of the first SNP-based linkage map using genotyping-by-sequencing and mapping of the <i>ms1</i> male-sterility gene in leaf chicory	160

Riassunto

I marcatori co-dominanti, tra cui i Microsatelliti (o SSR), sono strumenti molecolari ampiamente utilizzati nell'ambito della ricerca di base e applicata in specie di interesse alimentare. Tra le possibili applicazioni ricordiamo il loro impiego per studi di tracciabilità genetica di prodotti alimentari, per analisi di diversità genetica di varietà locali e identità genetica di varietà moderne e per il miglioramento genetico. Infatti gli SSR sono noti per essere altamente polimorfici e discriminanti, ben distribuiti all'interno del genoma, non influenzati da fattori ambientali, più efficienti e robusti dei marcatori fenotipici nelle analisi di diversità tra genotipi. Tuttavia, un'indagine condotta su 90 articoli scientifici basati sull'identificazione varietale delle specie economicamente più rilevanti in Italia, ha messo in luce la mancanza di un approccio comune tra gli autori in relazione alle strategie da utilizzare per questo tipo di studi. Inoltre lo studio ha evidenziato il bisogno improrogabile di stabilire procedure comuni riguardanti: i) i criteri da adottare per la scelta dei marcatori SSR ii) i parametri genetici più utili a questo scopo. Per dimostrare il potenziale di questa classe di marcatori, vengono presentati due casi studio. Il primo, che ha come oggetto Agordino, un'antica varietà locale veneta di orzo (*Hordeum vulgare* L.), ha permesso di enfatizzare la possibilità concreta di utilizzare i microsatelliti per la tracciabilità genetica di varietà locali ed, in particolare, di prodotti alimentari derivati. La caratterizzazione delle quattro principali varietà di mais (*Zea mays* L.) in Veneto -Sponcio, Marano, Biancoperla e Rosso Piave- attraverso marcatori SSR si è dimostrata invece estremamente utile per monitorare e prevenire fenomeni di erosione genetica, consentendo così di preservare la ricchezza genetica che le caratterizza, la loro identità fenotipica e i tratti qualitativi.

Nonostante l'interesse economico di alcune specie, non è così raro per i ricercatori doversi interfacciare con la totale mancanza di dati SSR e, più in generale, di informazioni genomiche. Finocchio (*Foeniculum vulgare* Mill., $2n=2x=22$), a tal proposito, rappresenta un esempio calzante. Per sopperire a questa carenza di dati, è stato condotto un sequenziamento su

piattaforma Illumina Hiseq 2500, permettendo così l'assemblaggio del prima bozza del genoma di finocchio in 300408 sequenze. La successiva annotazione ha consentito quindi di individuare e caratterizzare 103306 regioni altamente ripetute. Di queste, 40 scelte in modo casuale per il disegno di primer specifici, sono state testate e 14 sono state validate su una popolazione commerciale di 118 individui potenzialmente fruibili per lo sviluppo di ibridi F₁. Inoltre, il primo trascrittoma di foglia di finocchio è stato prodotto sovrapponendo due trascrittomi uno assemblato de novo e l'altro in silico, tramite allineamento sul genoma. 47775 dei 79263 trascritti totali sono stati annotati e 11853 risultano contenere una sequenza codificante completa. L'assemblaggio ha quindi consentito l'identificazione di loci coinvolti nella via biosintetica dei trans-anetolo, componente preponderante degli oli essenziali di finocchio e noto per le sue abilità nel ridurre dolori gastro-intestinali nonché per la sua attività antitrombotica e ipotensiva. Analisi dettagliate hanno infine messo in luce 1011 trascritti codificanti per fattori di trascrizione (FT), 6411 microsatelliti (EST-SSR), 3955 inserzioni/delezioni e 43237 polimorfismi a singolo nucleotide (SNP).

I marcatori di tipo SNP costituiscono un'altra classe di marcatori codominanti largamente sfruttati per la caratterizzazione di geni ad eredità Mendeliana e per l'analisi di poligeni o loci codificanti tratti quantitativi (QTL). Attraverso un approccio di genotipizzazione tramite sequenziamento (GBS) è stata costruita la prima mappa genetica in radicchio (*Cichorium intybus* L. subsp. *intybus* var. *foliosum*, 2n=2x=18) utilizzando una popolazione BC1 (ottenuta tramite tecniche di reincrocio) segregante 1:1 per il tratto "maschio sterilità". Questo studio ha permesso di localizzare finemente il gene nucleare della maschio sterilità *Cims1* all'interno del gruppo di associazione 9 e ha consentito l'identificazione di 4 SNP co-segreganti a 0 cM con il suddetto gene. Considerato che questa forma di maschio-sterilità, controllata da un singolo allele recessivo nucleare, è uno dei metodi più efficaci per produrre ibridi F₁, questi risultati saranno di estrema utilità per studi di miglioramento genetico.

General abstract

Co-dominant molecular markers, such as Microsatellites (or Simple Sequence Repeats, SSRs), are powerful tools for basic and applied research programs in crop plant species. Among the possible applications, they are frequently adopted for genetic traceability of food products, for assessing the genetic diversity of local varieties as well as the genetic identity of modern varieties, and also for marker-assisted breeding purposes. In fact, SSR markers are known to be highly polymorphic and discriminant, well distributed throughout the genome, not affected by environmental factors, more efficient and robust than phenotype-based field trials to detect and predict large numbers of distinct differences/traits among genotypes. However, a review of 90 original articles concerning the varietal characterization of some economically relevant crops in Italy, pointed out a lack of wider consensus among the authors regarding the strategy to design and to adopt for genotyping plant varieties with SSR markers. This study emphasized the urgent need to establish a common procedure concerning: i) the criteria adopted for selecting the marker loci and ii) the genetic parameters to be employed for varietal genotyping.

In order to demonstrate the potentials of these molecular markers, two case studies are presented. A study performed in Agordino, a very old local Venetian landrace of barley (*Hordeum vulgare* L.), stressed the concrete possibility to use SSR markers for genetic traceability of local varieties and, in particular, of their food derivatives. The genetic characterization of four main corn (*Zea mays* L.) landraces grown in Veneto (Italy), namely Sponcio, Marano, Biancoperla and Rosso Piave, by means of SSR markers, has shown great utility for monitoring and preventing further genetic erosion, thus preserving their gene pools, phenotypic identities and qualitative traits.

Despite the economic relevance of some crop species, it is common for researchers to deal with the complete lack of SSR data and, more in general, of genomic information. Fennel (*Foeniculum vulgare* Mill., $2n=2x=22$) represents a brilliant example. To overcome this shortage, an Illumina HiSeq 2500 sequencing was carried out in this species, enabling the assembly of the first genome

draft in 300,408 scaffolds. The subsequent annotation, permitted to detect and to characterize 103,306 SSR regions. Of these 40 were randomly chosen to design specific primer pairs, preliminary tested and 14 were successfully validated using a core collection of 118 fennel individuals potentially useful for F₁ hybrid development. Moreover, the first fennel leaf transcriptome was produced overlapping two transcriptomes, one assembled de novo, the other with an *in silico* genome-guided approach. A total of 47,775 out of the 79,263 assembled transcripts were annotated and, among them, 11,853 loci contained a putative full-length CDS. Detailed analysis revealed 1,011 transcripts encoding for transcription factors (TFs), 6,411 EST-SSRs, 43,237 SNPs and 3,955 In/Dels. Assembled transcripts were also used to conduct the identification of loci related to the t-anethole biosynthesis, the major component of the fennel essential oils, well-known for its capability in reducing mild spasmodic gastro-intestinal pains as well as for its antithrombotic and hypotensive activity. Finally, detailed analysis revealed 1,011 transcripts encoding for transcription factors (TFs), 6,411 EST-SSRs, 3,955 In/Dels and 43,237 SNPs.

Single nucleotide polymorphisms (SNPs) represent another class of co-dominant markers heavily exploited for the discovery of Mendelian inheritance genes and for the analysis of polygenes or QTLs (quantitative trait loci). Adopting a Genotyping By Sequencing (GBS) approach, the first SNP-based genetic linkage map of leaf chicory (*Cichorium intybus* L. subsp. *intybus* var. *foliosum*, 2n=2x=18) was built using a BC₁ population segregating 1:1 for the male sterility (ms) trait. This study enabled the genetic localization of the nuclear ms gene, termed *Cims1*, within linkage group 9 and the identification of four SNPs that proved to fully co-segregate with the target gene. Considering that this form of male-sterility, controlled by a single recessive nuclear gene, is one of the most effective methods to develop F₁ hybrids, our data will be exploitable for marker-assisted selection purposes.

Chapter I

Critical aspects on the use of microsatellite markers for assessing genetic identity of crop plant varieties and authenticity of their food derivatives

Abstract

A total of 90 original articles concerning the varietal characterization and identification by means of SSR analysis of the five most economically relevant crops in Italy (i.e. *Olea europaea* L., *Solanum lycopersicum* L., *Vitis vinifera* L., *Triticum* spp. and *Malus x domestica* Borkh.) have been selected and reviewed. Since the genetic traceability of processed products may result more complex, wine and olive oil have been considered too. Specifically, this chapter deals with three main aspects: i) the criteria adopted for the selection of the most appropriate number, type and distribution of SSR marker loci to be employed for varietal genotyping; ii) the use of genetic statistics and parameters for the evaluation of the discriminant ability and applicability of SSR marker loci; iii) how to make different experimental works on the same species standardized, reliable and comparable. What emerges from the studies here reviewed is a lack of wider consensus among the authors regarding the strategy to design and to adopt for genotyping plant varieties with SSR markers. This finding highlights the urgent need to establish a common procedure, especially for characterizing and preserving landraces, and for supporting its rediscovery and valorization locally.

Keywords: DNA genotyping, plant varieties, genetic traceability, food labeling, SSR

The Italian agriculture scenery and the utility of SSR markers to develop a reference method for genotyping plant varieties

The Food and Agriculture Organization (FAO) indices of agricultural production describe the relative level of the aggregate volume of agricultural production for each year in comparison with the base period 2004-2006 [1]. According to the most recent data available in The Food and Agriculture Organization Corporate Statistical Database, the gross value of the total Italian agricultural production was equal to \$ 41.9 billion, about € 32.7 billion [2]. It is worth noting that twenty products contribute to over 50% of gross production value (GPV), as shown in Table 1.

Table 1. GPV, registered cultivars, PDO and PGI products for the 20 most economically important crops in Italy.

Crop plants	Value [2] of agriculture production USD (10⁶)	Registered cultivars	PDO and PGI [4]
Olives (table and oil)	5064.24	644 [5]	3
Tomatoes	4753.00	445 [6]	3
Grapes (table and wine)	2770.60	638 [7]	3
Wheat (durum, common, spelt)	2558.23	489 [8]	0
Maize	2363.83	1739 [8]	0
Apples	1129.69	75 [9]	5
Oranges	990.80	n.a.	3
Potatoes	751.58	56 [8]	3
Rice, paddy	750.28	194 [8]	3
Peaches and Nectarines	570.69	311 [9]	4
Pumpkins	532.35	8 [6]	0
Pears	521.17	32 [9]	2
Mandarins, Clementines	431.27	n.a.	2
Artichokes	386.11	14 [6]	4
Carrots and Turnips	355.33	8 [6]	2
Cauliflowers and Broccoli	314.07	41 [6]	0
Beans	311.63	39 [6]	6
Lemons	267.16	n.a.	6
Onions	264.10	71 [6]	2
Hazelnuts (with shell)	247.36	25 [9]	3
Total	25333.49	4829	54

On average, each species is characterized by dozens or hundreds of cultivars and, as defined in Article 2 of the International Code of Nomenclature for Cultivated Plants, a “cultivar is an assemblage of plants that has been selected for a particular character or combination of characters,

that is distinct, uniform and stable in those characters, and that, when propagated by appropriate means, retains those characters”[3]. If some cultivars are virtually ubiquitous, some others are associated to specific geographical contexts and often provide the basis for the establishment of Protected Designation of Origin (PDO) and Protected Geographical Indication (PGI) products (Table 1).

It is not a coincidence that Italy, with its 268 brand products, including 106 PGI, 160 PDO and 2 Traditional Speciality Guaranteed (TSG) labels, is the European leader in terms of certified productions and that 20% of them arise from the 20 crops listed in Table 1. As a whole, the Italian certified products reach around 500 units, including two important derivatives such as olive oil and wine (Table 2).

Table 2. GPV, PDO and PGI products for Wine and Olive oil in Italy.

	Value [2] of agriculture production USD (10⁶)	PDO and PGI [10]
Wine	11603.83	403
Oil, olive, virgin	2126.78	41
Total	13730.61	444

It’s worth noting that the wine GPV (Table 2) is four times higher than the grape GPV and slightly less than half of the total GPV shown in Table 1, as demonstration that producing food derivatives could be more profitable than selling raw products.

One of the main problems that needs to be addressed is the lack of a uniform, complete and updated register of cultivars. For the cultivars of some species like cereals or vegetables are already available official registers provided by the Ministry of Agricultural, Food and Forestry Policies (MIPAAF, National register of agricultural varieties and National register of horticultural varieties). Concerning fruit trees, on the contrary, there is no a register yet, although Article 7 of the Italian Legislative Decree no. 124/2010 has established a “National Register of fruit trees varieties”[11]. For this reason, the inventory of cultivars of some fruit species is still ongoing and there is a total lack of official data for some of them (see for instance orange, lemon and mandarin, Table 1).

Moreover, for species of particular interest exist registers apart (see for example *Olea europaea* L. and *Vitis vinifera*, L.).

In the past, cultivars have been extensively characterized by morphological traits, including plant, leaf, fruit and seed characteristics. Since objectivity is crucial to perform an accurate morphological typing, it is constraining to use exclusively morphological descriptors for plant cultivars, especially because most of the morphological traits are influenced by environmental factors. Several cases of misidentification owing to classifications carried out only employing morphological traits, are reported in the scientific literature for a wide range of vegetal crops [12–14] and fruit tree [15–18]. Moreover, the uneven distribution, simultaneous cultivation of local varieties, ambiguous names, continuous interchange of plant materials among varieties and/or farmers of different regions and countries, possibility of the cultivation of varietal clones, and uncertainty of varietal certification in nurseries have complicated identification of genotypes [19–21]. At the same time, cultivar and clone identity is also very important for protecting plant breeders' rights not only for commercial seeds but also for processed materials and food derivatives, especially for the final consumers' safeguard. Another important aspect to highlight is the need to ensure that each specific variety grown by farmers and its food product bought by consumers is the one declared on the label. This is especially true if the product is sold in a processed or transformed form (thus difficult to recognize phenotypically) and/or if the product is subjected to a form of certification (PDO or PGI). In a modern market, it is crucial being able to identify agricultural products and foodstuffs by means of reliable traceability systems, including genetic molecular markers.

The method of DNA genotyping based on microsatellite markers represents an efficient, reliable and suitable technique that is able to complement the information provided by morphological traits and that has been extensively used for the characterization of plant varieties [22–24] and the certification of food products [25–27].

Microsatellites (or simple sequence repeats, SSRs) are PCR-based molecular markers valued for their abundant and uniform genome coverage, high levels of polymorphism information content as a consequence of their marked mutation rates, and other valuable qualities such codominant inheritance of DNA amplicons/alleles and request of little amount of DNA for the amplifications [28]. A unique pair of primers defines each SSR marker locus, as a consequence the molecular information exchange among laboratories is easy and allows individuals to be uniquely genotyped in a reproducible way [29].

SSR markers have been shown repeatedly as being one of the most powerful marker methodologies for genetic studies in many crop species. In fact, since they are multiallelic chromosome-specific and well-distributed in the genome, microsatellite markers have already been used for mapping genes with Mendelian inheritance [30], for identifying quantitative trait loci (QTLs, [31]) and for molecular marker-assisted selection [32]. In many species, microsatellite markers have also been used for ascertaining the genetic purity of seed lots [33], as well as to assess the capability to protect the intellectual property of plant varieties [34]. These markers are also largely used for assessing the genetic diversity and relationships among populations and lines, and for identifying crop varieties.

The advantages of SSRs over single-nucleotide polymorphisms (SNPs), another co-dominant marker systems increasingly exploited in breeding programs, include relative ease of transfer between closely related species [35,36] and high allelic diversity [37,38]. On the contrary, SSRs when compared to SNPs have some limits: the development phase is quite long and expensive for multi-locus assays and the throughput is relatively low because of drawbacks for automation and output data management. Recently, progresses in the development of multi-locus assays have been made in several directions, suggesting that SSR markers still remain relevant molecular tools at least for specific applications and genetic studies [39]. In fact, PCR-based SSR genotyping has rapidly evolved in plants, and methods for the simultaneous amplification of multiple marker loci coupled to semi-automated detection systems have been developed [40]. The identification and

selection of SSR markers have become cheaper and faster due to the emergence of next-generation sequencing technologies means. Moreover the possibility to multiplexing specific combinations of microsatellite markers has become much easier and the availability of capillary electrophoresis equipment relying on automated laser-induced fluorescence DNA technology has facilitated the adoption and exploitation of this methodology in applied breeding programs [41–43].

Genotypic characterization through SSR loci analysis represents a molecular tool applicable to all species and able to support the phenotypic observation in order to characterize and describe a cultivated variety as well as to define its uniformity, distinctiveness and stability (DUS testing). At the same time, SSR markers are largely used for the genetic identification of varieties and the authentication and traceability of their foodstuffs [44–46].

The main goal of this work is to provide an updated and detailed description of the applications of SSR markers for varietal characterization and identification, reviewing the state of the art of genotyping in the most economically relevant Italian crop plants and food products: *Olea europaea* L., *Solanum lycopersicum* L., *Vitis vinifera* L., *Triticum* spp. and *Malus x domestica* Borkh., wine and olive oil. In this respect, the chapter aims to assess the real achievements of different genotyping analyses, to evaluate the strengths and limitations according to applied research studies, and to emphasize the striking lack of data related to the applications of SSR technology. Through the careful investigation and evaluation of a large number of scientific papers, our review highlights some critical aspects on the use of microsatellite markers and formulates recommendations for standardizing the strategies and methods for ascertaining the genetic identity of plant varieties and for achieving the genetic traceability of their food derivatives. Here we focus on three main aspects: i) how to choose and use SSR markers; ii) which parameters/indices calculate for the genetic characterization of plant materials; iii) assess a standardized way to make SSR data from different works on the same species comparable.

Applications of SSR markers for the genetic characterization of crop plant varieties

Some of the most economically important crops in Italy have been chosen for this study and the search has been focused on their varietal characterization through SSR analysis. In particular, olive (*Olea europaea* L.), grape (*Vitis vinifera* L.) and apple (*Malus × domestica* Borkh.) were reviewed among the fruit trees whereas wheat (*Triticum* spp.) and tomato (*Solanum lycopersicum* L.) were selected as representative of cereals and vegetables, respectively. A large number of commercial cultivars are available for each of these species and the annual Italian GPV for these crops is about 18 billion euro [2]. Moreover, scientific articles dealing with the genetic identification in wines and olive oils were also evaluated because these two derivatives contribute to the annual Italian GPV for another 15 billion euro [2].

Although passport data, morphological and agronomical descriptors have been collected, data are not informative enough to assess the numerous cases of misidentification, mislabelling, homonymies and synonymies as well as voluntary or accidental frauds [47]. With regard to this, several research groups characterized and identified cultivars using SSR markers (Table 3).

Table 3. Crops and derivatives reviewed.

Crops	References
Olive (<i>Olea europaea</i> L.)	[23,25,48–63]
Tomato (<i>Solanum lycopersicum</i> L.)	[27,44,64–70]
Grape (<i>Vitis vinifera</i> L.)	[15,19,24,71–89]
Wheat (<i>Triticum</i> spp.)	[22,26,90–98]
Apple (<i>Malus x domestica</i> Borkh.)	[99–111]
Derivatives	
Wine	[45,112–116]
Olive oil	[46,58,117–125]

Article searches were performed using the three most popular sources of scientific information: Scopus, Web of Science and Google Scholar, while PubMed was excluded from the queried datasets because it focuses mainly on medicine and biomedical sciences and also because Google scholar already includes its index [126]. A total of 90 articles based on SSR genotyping analysis were selected from the international literature in the last 15 years, covering all the plant

species/food products taken as reference list. Only articles dating from 2000 to now were reviewed assuming that researches published earlier would have lost their steering effects on the activities of plant DNA genotyping, given that the development of new and large marker datasets, and technologically advanced and automated protocols has been very fast in the last 15 years.

What number and how to select a panel of SSR marker loci according to their linkage map position and polymorphism information content

More than 800 SSR markers have been developed in apple (*Malus x domestica* Borkh., $2n=2x=34$) and nearly all of them have been mapped on a consensus map produced starting from 5 different genetic maps [127]. These markers are distributed across all 17 linkage groups, with an average of 49 microsatellites per linkage group. Moreover, the Genome database for Rosaceae [128] is a long standing community database resource providing hundreds of microsatellite loci, in most cases accompanied by a wealth of information about map position, repeat motifs, primers, PCR conditions, amplicon length and publication source. A discriminatory set of markers should ensure the uniform distribution across the genome of the microsatellite loci to represent adequately each linkage group and, thus, the genome in its entirety [91]. In fact assessing the genetic diversity by focusing only on restricted regions of the genome may threaten to distort results. Nevertheless, neglecting the most ambitious study on *Malus x domestica* Borkh. carried out by Patocchi et al. [105] using an extremely high number of SSR markers (82), the number of selected and analyzed genomic loci varies from 4 to 19 with an average value of 12 ± 6 SSR markers, less than a microsatellite locus per linkage group. Extending this reasoning also to the other crops reviewed, the emerging output is often the same: for all the plant species very detailed genetic maps are available [129–132] as well as dedicated databases for SSR markers (Table 4)

Table 4. Information on the five species analyzed in this book chapter, including genome size, ploidy, available SSR database and number of microsatellite regions included, average number of SSR employed in the articles reviewed, number of cultivars and microsatellite used as reference.

Species	Genome size (Gb)	Ploidy	SSR available (SSR database)	SSR employed (mean±st.dev)	N. of reference cultivars	N. of reference SSRs
<i>Olea europaea</i> L.	1.42–2.28 [133]	2n=2x=46	12 (OLEA Database) [134]	11±5	21 [53], 17 [52]	11 [53], 8 [52]
<i>Solanum Lycopersicum</i> L.	0.90-0.95 [132]	2n=2x=24	146,602 (Tomato microsatellite database) [135] 66,823 (Tomato genomic resources database) [136] 21,100 (Tomato - Kazusa Marker Database) [137]	14±7	n.a.	n.a.
<i>Vitis vinifera</i> L.	0.48 [129]	2n=2x=38	56 (Grape Microsatellite Collection) [138] 443 (Italian Vitis Database) [140] 6 (The European Vitis Database) [141]	15±11	49 [139]	6 [139], 38 [74]
<i>Triticum</i> spp	12.3-13.00 (<i>T. durum</i> Desf) [142] 16.50-17.00 (<i>T. aestivum</i> L.) [142]	2n=4x=28 2n=6x=42	588 (Wheat Microsatellite Consortium) [143]	18±3 21±6	n.a.	46 [144]
<i>Malus x domestica</i> Borkh.	0.75 [145]	2n=2x=34	664 (HiDRAS SSR database) [146] 2,449 (Genome database for Rosaceae) [128]	12±6	7 [147]	12 [147], 15 [108]

Olea europaea L. ($2n=2x=46$) includes 23 chromosome pairs and the average number of microsatellite markers used in the reviewed articles is 11 ± 5 , much less than a microsatellite locus per linkage group. The same is true also for *Vitis vinifera* L. ($2n=2x=38$) in which the average number of microsatellite markers explored for genotyping cultivars is 15 ± 11 in spite of the 19 chromosome pairs of this species. Even the varietal identification of their respective derivatives (olive oil and wine) has been accomplished by exploring, on average, 8 ± 3 and 10 ± 4 SSR markers, respectively. On the contrary, in wheat the varieties of both *Triticum durum* Desf. ($2n=4x=28$) and *Triticum aestivum* L. ($2n=6x=42$) have been characterized by means of genotyping with SSR markers analyzing, on average, 18 ± 3 and 21 ± 6 microsatellite loci respectively, that is more than one microsatellite per linkage group. This latter choice is perhaps associated to the high complexity and large size of the *Triticum aestivum* L. genome, approximately equal to 17 Gb/1C [148]. In fact, for a correct representation of the entire genome, not only the number of homologous chromosomes but also their size (i.e. total amount of DNA) should be considered when choosing the optimal panel of microsatellite loci to be investigated. Finally, in tomato (*Solanum lycopersicum* L., $2n=2x=24$) the average number of SSR markers employed for genotyping varieties is 14 ± 7 (Table 4).

Only few studies [65,74,96,106] evaluated the position within linkage groups of the microsatellites selected: the choice often falls on SSR markers with unknown or not specified position or mapped on few chromosomes, thus resulting in a poor representation of the entire genome. In this regard, the results from Cipriani et al. [74] and van Treuren et al. [106] represent a good model for the choice of molecular markers to investigate the genetic diversity in germplasm collections and to solve synonymy/homonymy cases as well as paternity and kinship issues. The former group selected microsatellite sequences from scaffolds anchored to the 19 linkage groups of *Vitis vinifera* L. with the aim of analyzing 38 well-distributed SSR markers, ideally two loci for each linkage group, whereas the latter group also considered the specific map position genetic and genetic association with traits of agricultural interest.

Two important issues must be pointed out. The number of SSRs to employ should be also evaluated according to the type of analysis. For example, the EU-Project Genres CT96 No81 [139] selected six highly discriminating microsatellites, thus less than one marker per linkage group, that could be sufficient to differentiate among hundreds of grape cultivars. The same microsatellite set could be very inadequate to discriminate among clones. Moreover, it is worth noting that, in some cases, increasing the number of marker loci does not necessarily mean improving the resolution of cultivar characterization and identification. For example, Baric et al. [107] reported that extending the set of microsatellite markers to 48, from an initial analysis based on 14 SSR loci, it was impossible to improve the genetic discrimination among the 28 accessions of *Malus x domestica* Borkh. analyzed. Connected to the distribution and position of the microsatellite loci within a genome, there is also the possibility to choose between genomic SSR (gSSR) and EST-derived SSR (EST-SSR). Generally EST-SSR markers are less polymorphic than genomic SSR ones, as reported for *Triticum* spp. [93,95] and *Solanum lycopersicum* L. [68], being the formers found in selectively more constrained regions of the genome. Of particular interest is the comparison of Leigh et al. [93] between sets of 20 EST-SSR and 12 genomic SSR markers in terms of discrimination ability among 66 varieties of *Triticum* spp. The results indicate that the panel of EST-derived SSR markers used is slightly less efficient at discriminating between hexaploid *Triticum aestivum* L. varieties compared to the second panel of genomic SSR markers. EST-SSR markers also have the disadvantage that amplicon sizes can differ from expectations, as a consequence of the undetected presence of introns in flanking regions [39]. Nevertheless, these findings support the possibility that EST-SSR markers could in the near future complement and outnumber the genomic SSR markers. In fact, EST-SSR markers should have some important advantages over genomic SSR markers. In particular, they are easily obtained by bioinformatic querying of EST databases while the development phase of genomic SSR markers is quite long and expensive; EST-SSR markers could be functionally more informative than genomic SSR markers because associated with transcribed regions of the genome,

thus reflecting the genetic diversity inside or adjacent to the genes [149]. Moreover, the rate at which SSR flanking regions evolve is lower in expressed than non-expressed sequences and the primers designed on these sequences are more likely to be conserved across species, thus resulting in high levels of SSR transferability [150]. A suitable combination of EST-SSR and genomic-SSR markers could be optimal for distinctiveness, uniformity and stability testing applications for crop plant varieties [93]. Overall, the vast majority of studies are based on genomic SSR markers, and only three articles out of 90 take into account the possibility of employing EST-SSR markers.

In terms of location, nuclear SSR (nSSR) markers are largely used and more exploited than plastidial and mitochondrial SSR (cpSSR and mtSSR, respectively) markers. First, the development phase of extra nuclear SSR markers is complicated: high purity chloroplast or mitochondrial DNA is typically very hard to extract due to nuclear DNA contaminations [151]. Moreover, Wolfe et al. [152] have shown that comparing nuclear, chloroplast and mitochondrial genomes, the frequency of chloroplast genome gene silencing and replacement was half that of the nuclear genome, and three times that of the mitochondrial genome, indicating that the evolution of mitochondrial genome has been slower and implicating lower levels of polymorphism. Nevertheless, the use of markers belonging to mitochondrial or chloroplast sequences may be useful due to their haploid nature, relative abundance and stability in comparison with nuclear sequences. For instance, Borgo et al. [153] suggested that the circular form increases stability and resistance against heat disintegration. Boccacci et al. [113] analyzed musts and wine samples using a set of 9 nSSR and 7 cpSSR markers in order to identify cultivars. Findings from these studies confirm a low level of polymorphism for the extra nuclear markers due to their lower frequency of mutation. Also Baleiras-Couto and Eiras-Dias [45] and Pérez-Iménez et al. [125] have exploited this kind of SSR markers, with similar results.

The choice of the number of SSR loci usually depends on their polymorphism degree. With some exceptions for which this information is not available, the average number of marker alleles per

SSR locus is equal to 7.1 for *Olea europaea* L., 3.5 for *Solanum lycopersicum* L., 8.2 for *Vitis vinifera* L., 6.9 for *Triticum* spp., 9.4 for *Malus x domestica* Borkh., 6.5 for olive oil and 5.2 for wine. Both EST-SSR and cpSSR were found to be less polymorphic, with a low average number of alleles per locus, than genomic SSR markers [45,68,93,113,125]. The polymorphism degree may depend on several factors, including the SSR motif length and the SSR localization on coding or not-coding regions.

In order to estimate the level of genetic diversity detected by each microsatellite, marker frequencies are widely used to estimate the polymorphism information content (PIC, Table 5) values, according to the methods of Botstein et al. [154]. The authors reported the following formula for the calculation of the PIC value of an n -marker allele:

$$PIC=1-\sum_{i=1}^n p_i^2-\sum_{i=1}^{n-1} \sum_{j=i+1}^n 2p_i^2 p_j^2 \quad (1)$$

where p_i and p_j are the population frequencies of the i^{th} and j^{th} marker alleles, respectively. A $PIC > 0.5$ is considered as being highly informative marker, while $0.5 > PIC > 0.25$ is an informative marker and PIC is 0.25 a slightly informative marker. As reported by Nagy et al. [155], PIC can be defined as the probability that the marker genotype of a given offspring will allow deduction, in the absence of crossing-over, of which of the two marker alleles of the affected parents it received. In other words, this parameter is a modification of the heterozygosity measure that subtracts from the H value an additional probability that an individual in a linkage analysis does not contribute information to the study. On this aspect, there is not full agreement among the authors. Some studies on olive oil [58,122] and *Malus x domestica* Borkh. [101,103], referring to Anderson et al. [156], contend that the occurrence of rare marker alleles has less impact than common marker alleles on the PIC estimates and consider that this index can be assimilated to the expected heterozygosity (H_e), calculated by the following simplified formula:

$$PIC=1-(\sum_{i=1}^n p_i^2) \quad (2)$$

where p_i is the population frequency of the i^{th} marker allele.

In addition to the PIC value, calculated taking into account allelic frequencies, there are several indexes focusing on genotype frequencies. For example, as reported by Aranzana et al. [157], other two important indexes that should be evaluated are the power of discrimination (usually PD) -or diversity index (D), as reported by Zulini et al. [71] and Martínez et al. [24]- and the Confusion Probability (C). The first one provides an estimate of the probability that two randomly sampled accessions of the study would be differentiated by their marker allele profiles:

$$PD=1-\sum_{i=1}^n p_i^2 \quad (3)$$

where p_i is the frequency of the i^{th} marker genotype. As already described for the PIC, among the authors there are different interpretations and procedures to calculate the PD index. Pasqualone et al. [25] in their study on *Olea europaea* L. genotyping reported that “the power of discrimination, sometimes referred to as polymorphism information content, or diversity index, was calculated [...]”, assuming in this way that PD and PIC correspond to the same parameter.

The confusion probability (C) index, also defined as the combined power of discrimination overall loci [23] is the probability that any two cultivars are identical in their genotypes at all SSR loci by chance alone and it depends on PD. It can be estimated as follows:

$$C=\prod_{i=1}^n (1-PD_i) \quad (4)$$

where PD_i is the power of discrimination value of the i^{th} locus. Notwithstanding its informativeness, only 3 articles of the 90 reviewed take into account this value (Table 5). Martínez et al. [24] in their

attempt to assess the genetic diversity of *Vitis vinifera* L. varieties calculated the power of discrimination index as follows:

$$PD=1-C \quad \text{being} \quad C=\sum_{i=1}^n p_i^2 \quad (5)$$

where p_i is the frequency of different marker genotypes for a given locus. In this case, C is the probability of coincidence, corresponding to the probability that two varieties match by chance at one locus.

Twenty-one articles, mainly focused on the species *Vitis vinifera* L. and oil from *Olea europaea* L., report also the probability of identity (PI) index of each single SSR marker locus either in addition or in substitution of PD value (Table 5). This index can be estimated as follows:

$$PI=\sum_i (p_i)^4 + \sum_i \sum_j (2p_i p_j)^2 \quad (6)$$

where p_i and p_j are the frequencies of i^{th} and j^{th} marker alleles, respectively. It represents the probability that two individuals drawn at random from a population will have the same genotype at one marker locus. For example, Vietina et al. [122] and Corrado et al. [58] in their studies regarding the genetic traceability of monovarietal olive oils, refer to this value in order to determine the efficacy of the SSR marker pool to discriminate among the cultivars. Martínez et al. [24] adopted the following formula to calculate the same value:

$$PI=\sum_i (p_i)^4 - \sum_i \sum_j (2p_i p_j)^2 \quad (7)$$

Equally interesting is the total probability of identity (PI_t) that represents a compound probability defined as the probability of two cultivars sharing the same marker genotype by chance and calculated as follows:

$$PI_t = \prod_{i=1}^n PI_i \quad (8)$$

where PI_i is the probability of identity value of the i^{th} marker locus.

Finally, Qanbari et al. [158] reported that PD and PI are complementary parameters:

$$PD=1-PI \quad (9)$$

The use of standardized parameters is essential to make SSR data comparable across species and laboratories and it can be especially beneficial for the preliminary evaluation of the discriminant ability and applicability of SSR marker loci.

Table 5. Summary information on the main parameters assessed by the 90 papers reviewed.

Index	Full name	Formula	Definition	N. of papers account for it
PIC*	Polymorphism Information Content	$1 - \sum_{i=1}^n p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^n 2p_i^2 p_j^2$	Probability that the marker genotype of a given offspring will allow deduction, in the absence of crossing-over, of which of the two marker alleles of the affected parents it received [155]	36
PD**	Power of Discrimination	$1 - \sum_{i=1}^n p_i^2$	Probability that two randomly sampled accessions would be differentiated by their marker allele profiles [157]	14
C**	Confusion probability	$\prod_{i=1}^n (1 - PD_i)$	Probability that any two individuals are identical in their genotypes at all SSR loci by chance alone [157]	3
PI*	Probability of Identity	$\sum_i (p_i)^4 + \sum_i \sum_j (2p_i p_j)^2$	Probability that two individuals drawn at random from a population will have the same genotype at one marker locus [122]	21
PI _t *	Total probability of identity	$\prod_{i=1}^n PI_i$	Probability of two individuals sharing the same marker genotype by chance [122]	2

* pi and pj are the frequencies of the ith and jth marker alleles.

** pi is the frequency of the ith marker genotype.

The choice of the best microsatellite motifs and the problem of the null alleles

Microsatellite repeat units typically vary from one to six bases. Shortest motifs (mono- or dinucleotide repeats) usually have a high number of alleles [74] and they allow packing more loci on a given separation system, resulting in larger multiplexes. However, this kind of SSR motifs can be difficult to assay accurately. It is very common to observe a stuttering in terms of multiple bands or peaks, a phenomenon commonly caused by slippage of the DNA polymerase, but the main problem arises when there is a difference of one or two base-pairs between marker alleles: in case of homozygous loci the electrophoretic analysis results in one main band or peak, but with heterozygous loci very often one of the two marker alleles is masked by the stutter. SSR markers containing trinucleotide or higher order repeats usually eliminates this technical problem because target sequences appear to be significantly less prone to slippage [52]. Nevertheless, microsatellite loci with long motifs are known to be less polymorphic and, in some cases, due to lack of stutter bands or peaks, is not always possible to distinguish SSR amplicons from other aspecific PCR products and it may lead to an overestimation of the level of polymorphism of these loci [159].

Among the 90 studies we surveyed, only 25 of them specify the length of the SSR motifs employed and very few justifies the choice. Cipriani et al. [80] performed two distinct molecular analyses on the same set of cultivars, using the genetic profiles obtained from the two sets of microsatellites, the di-nucleotide repeats from one side, and the tri-, tetra- and penta-nucleotide repeats from the other, with the aim of comparing their performance in the discrimination of the genotypes analyzed. Both microsatellite data sets produced identical consensus tree topology, but the authors underlined that di-nucleotide SSR markers, scored a higher number of alleles per locus, and consequently, a potentially higher power for identifying and distinguishing closely related genotypes. On the other hand, the microsatellite dataset based on tri-, tetra- and penta-nucleotide SSR markers proved to have the advantage of ease in scorability, while maintaining a very high power of discrimination for successful genotyping of the *Vitis vinifera* L. cultivars.

Microsatellites have also been classified according to the type of repeat sequence as perfect or imperfect, according to the occurrence of simple or uneven repeats, respectively [160]. The preference should be given to perfect motifs because using imperfect ones there is no more equivalency between fragment length and amplicon sequence, and hence several sequences can correspond to a given length variant [39]. This is the reason why only four studies employed imperfect SSRs among the 25 ones specifying the motifs.

The occurrence of null alleles is something to avoid when using SSR markers for genotyping plant materials. A microsatellite null allele is any marker allele at a genomic locus that consistently fails to amplify by the polymerase chain reaction, resulting in the lack of detectable amplicons. Lack of amplified fragments could preclude the detection of heterozygous loci, which would be computed as homozygotes. In the same way, null alleles at homozygous loci are characterized by a completely lack of amplification with the consequent production of missing data. On the whole, null alleles may interfere with the genetic identification of cultivars, by wrongly reducing the genetic diversity among accessions [149]. In the 90 studies surveyed, only 38 of them estimated the probability of null alleles, mainly using the formula of Brookfield [161]:

$$r = \frac{H_e - H_o}{1 + H_e} \quad (10)$$

being H_e the expected heterozygosity and H_o the observed heterozygosity.

Comparisons across studies of SSR-based genotyping: reference marker sets and reference plant varieties

In most cases, it is impossible to make valid comparisons across studies on the same species since different sets of SSR loci are used in different laboratories [162]. For some species the choice of microsatellites begins to be fairly uniform (Table 4). For instance, almost all of the studies aimed to genotype *Olea europaea* L. cultivars make use of SSR markers belonging to four main datasets

developed by Sefc et al. [163], Carriero et al. [164] Cipriani et al. [165], De La Rosa et al. [166]. Based on these studies, two informal universal sets of SSR markers were proposed for genotyping *Olea europaea* L. cultivars by Doveri et al. [52] and Baldoni et al [53]. Cipriani et al. [74] suggested a list of 38 markers with excellent quality of peaks, high power of discrimination, and uniform genome distribution (1–3 markers/chromosome) for genotyping *Vitis vinifera* L. cultivars. Li et al. [144] assembled a reference kit of SSR markers for genetic analysis in *Triticum* spp. that comprises 46 microsatellites. Moriya et al. [108] developed a set of SSR markers for genotyping *Malus x domestica* Borkh. cultivars that includes 15 microsatellites. Not only independent research works, but also some international programs and projects attempted to pursue this goal. The European Cooperative Programme for Plant Genetic Resources (ECPGR) has recommended a new set of 12 SSR marker loci distributed in different linkage groups of the *Malus x domestica* Borkh. genome, organized in three multiplexes and designed for a four-dye system [147]. Comparable considerations have been presented within two projects focused on the grapevine genetic resources conservation and characterization (EU-project GENRES CT96 No 81, [139]) and on the Traceability of Origin and Authenticity of Olive Oil (Oliv-Track, [167]). It is worth noting that, to the best of our knowledge, for *Solanum Lycopersicum* L. no SSR set of reference has been proposed yet.

Unfortunately, establishing a reference set of microsatellite markers to use in each analysis for a given species it is not sufficient to ensure the comparability among different studies and the reproducibility among different laboratories. Some tests have been carried out in order to investigate the reproducibility of SSR data produced by different laboratories under varying local conditions. Four different laboratories performed independent marker analyses on a common set of 21 DNA samples of *Olea europaea* L. cultivars and with the same set of SSR markers, using different DNA polymerase enzymes, PCR cycling conditions, amplicon separation and visualization methods [53]. The results are not encouraging. Many cases of allele drop out and discrepancies in

allele length, up to five nucleotides for identical microsatellite loci were recorded. This finding is probably attributable to a combination of different equipment, different sequencers and different internal ladders, which may have affected the relative mobility estimates leading to non-comparable electropherograms. Similar results have been achieved from ten laboratories distributed in seven countries that analyzed the same 46 *Vitis vinifera*, L. cultivars at the same 6 SSR loci [72].

One of the main discoveries is that the specific microsatellite sequence dramatically influences the efficiency of analysis. Marmiroli et al. [168] showed that the repeatability of results among different laboratories was good enough for some microsatellites, but rather low for others confirming that the choice of SSR loci and of their primers is crucial for an efficient analysis.

Despite all the precautions and the establishment of a reference set of SSR markers, some residual variation in laboratory equipment and procedures cannot be completely avoided, and representative reference material with many different alleles should be adopted by all laboratories involved in a genotyping program for a given species [162]. For this purpose, 21 out of 90 studies included reference cultivars, promoting new ones or exploiting cultivars already used as reference in previous works. Independent researches and international institutions are trying to find an agreement filling lists of reference accessions in order to prevent that each group uses its own reference cultivars and to standardize all works performed on these species. For example, the ECPGR has chosen eight *Malus x domestica* Borkh. cultivars as reference set for this species [147]. Baldoni et al. [53] and Doveri et al. [52] proposed two different lists of reference cultivars for *Olea europaea* L. (Table 4).

Even if this approach is fully applicable also to the crop derivatives here taken into account (olive oil and wine), there are some additional aspects that must be considered when talking about processed products. First, sometimes it is very difficult to make SSR marker analyses on food products and beverages because of the low DNA quantity and the lack of DNA integrity. For example, Baleiras-Couto and Eiras-Dias [45] reported their difficulties to investigate wines after

about eight months of fermentation, as well as Recupero et al. [115] highlighted technical problems during the isolation of genomic DNA from Nebbiolo wine. Nevertheless, both of them managed to characterize must. For olive oil, Martins-Lopes et al [119] as well as for Vietina et al. [122], took advantage from extraction methods able to give good yield of genomic DNA and PCR amplificability. Is therefore evident how an optimized DNA extraction method is also a crucial step to carry out a reliable study on the applicability of molecular markers for identifying the varietal origin or assessing the varietal composition of crop plant derivatives.

It is not trivial considering the match between genetic profiles of crop plants and their derivatives. In this regard, there are some contrasting points of view. In the review of Agrimonti et al. [169] is reported that several authors [e.g. 46,118,120] have noticed a satisfying conformity between olive oil and leaf profiles with SSR markers. On the contrary, Doveri et al. [117] have proposed a cautionary note about the use of SSR markers, stressing the non-perfect concordance between the molecular genetic profiles of the olive oil and the original leaf sample. Furthermore, it is necessary to underline the extreme difficulty in characterizing multi-varietal derivatives through SSR analysis. Most of the Italian PDO wines and olive oils are produced blending two or more cultivars in percentages strictly defined in the production regulation. In these cases, each SSR locus is represented by the combination of the marker alleles of each variety. For examples, Baleiras-Couto and Eiras-Dias [45], after having analyzed with six SSR markers different di-varietal musts at different percentages, reported results that confirm the complexity and difficulty of assessing multiple genotypes.

Conclusions

The genetic characterization of plant varieties by means of multi-locus genotyping through SSR markers in the main crop species is still not based on standardized protocols making difficult the acquisition of reproducible and transferable datasets. What emerges from the analysis of the

literature is a lack of wider consensus among the authors regarding the strategy to design and to adopt for genotyping plant varieties with SSR markers. This finding highlights the urgent need to establish a common procedure.

Some conclusions of general validity can be drawn on the basis of the articles here reviewed. First of all, it is quite difficult to define exactly the ideal number of microsatellite loci to assay. Usually, the number of SSR markers depends on the type and goal of the analysis. If the purpose is merely to distinguish among two or more cultivars (i.e. individual genotypes), it is possible to adopt an “as simple as possible strategy”. For example, a novel approach called the Cultivar Identification Diagram (CID) strategy has been recently developed. This method was designed so that, at each step, a polymorphic marker generated from each PCR analysis directly allows the separation of cultivar samples [109]. In this specific study, eight is considered the minimum number of SSR markers necessary to distinguish 60 cultivars in *Malus x domestica* Borkh.. Supposedly, the number of SSR markers could depend on the number of cultivars to distinguish, on their relationship and on the polymorphic degree of each marker locus. In this regard, we suggest AMaCAID [170] and UPIC [171], two very interesting tools that able the investigation of the minimum number of markers required to distinguish a specific number of accessions and, thus, the identification of the best marker combination that maximizes the genetic information.

When the purpose is to genetically characterize a cultivar in order to fulfil the requirements of a varietal register that could include hundreds or thousands of different varieties, the selection of SSR markers should be oriented to an exhaustive representation of the genome as whole. This is the reason why different authors consider one or two microsatellite for each linkage group the minimum number required to reconstruct a reliable and selectable genotype for a given plant accession. For instance, Cipriani et al. [74] implemented an efficient method for *Vitis vinifera* L. fingerprinting using a set of 38 microsatellite marker loci scattered throughout the genome. In particular, two SSR loci were carefully chosen, on average, for each linkage group, selecting the

best ones in terms of polymorphism information content (PIC), and power of discrimination (PD, Figure 1).

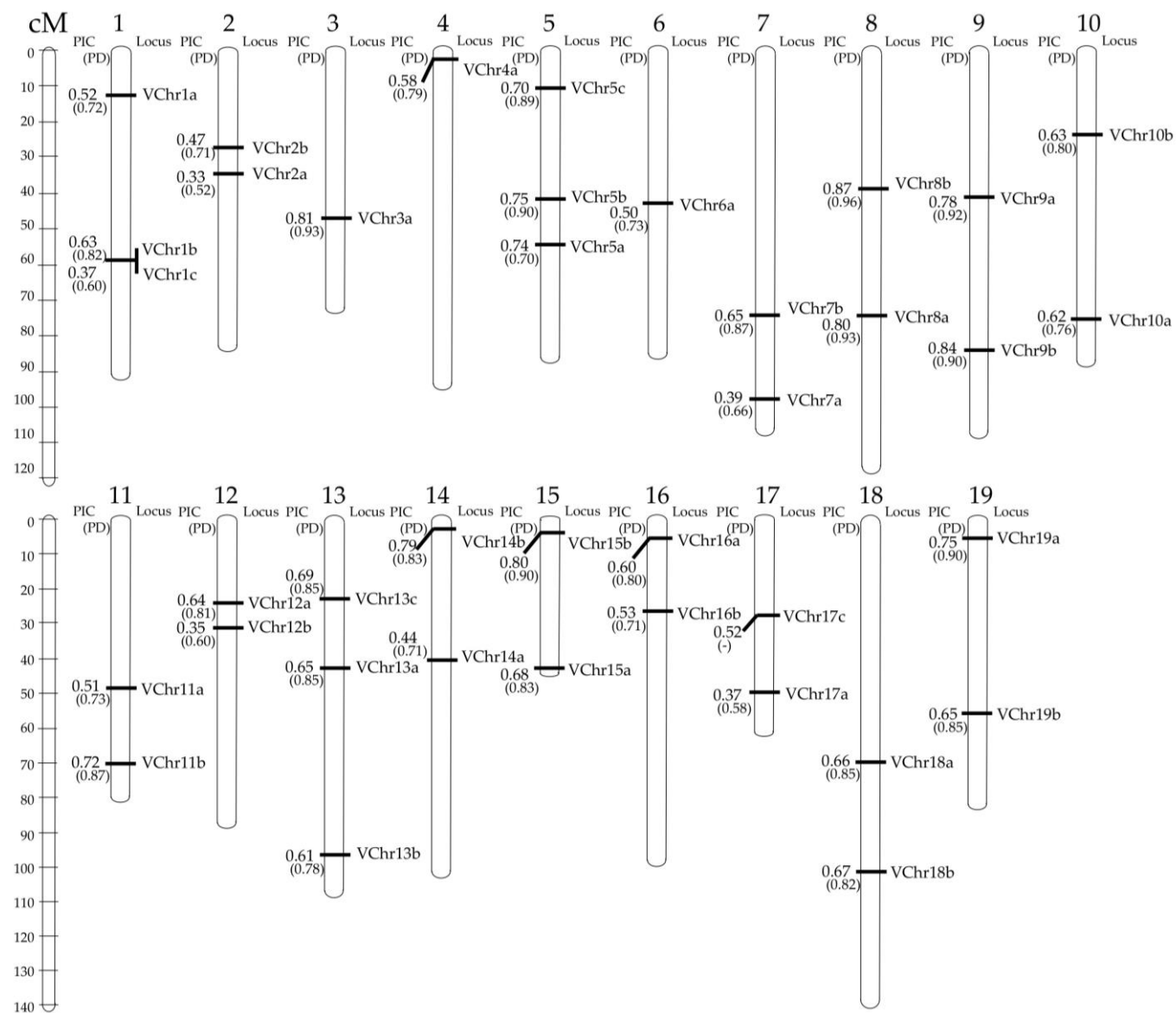


Figure 1. Schematic representation of the 19 basic linkage groups of *Vitis vinifera* L. with indication of the 38 mapped marker loci chosen on the basis of their discriminant informativeness. In addition to the marker name, each locus reports the individual power of discrimination (PD) and the polymorphism information content (PIC). Figure modified from Cipriani et al. [74].

It is worth noting that despite some international programs and projects attempted to establish reference SSR set, there is still a lack of wider consensus. For instance, in 2003, the partners of the EU-project Genres CT96 No81 [139] agreed on the utilization of six highly polymorphic SSR-markers for the identification of *Vitis vinifera* L. cultivars, but, since then, several studies continue to be performing using a higher number of markers [74,76,78,84,86]. As reported by Cipriani et al. [74], grape varieties selected in Western Europe, which account for most of the worldwide production of wine, likely have extensive coancestry that is common origin from the hybridisation of a few ancestors. Because of this, using too few markers for fingerprinting could hamper the discrimination of sibling varieties. For this reason, they recommend using at least 19 markers (among the 38 markers employed in their work). In general, for the selection of the panel of SSR markers, the following criteria should be followed. Based on previous works, the SSR marker loci with the highest number of marker alleles and the highest PIC and PD scores should have the priority. In addition, the position of the SSR markers across the genome, as mapped in different linkage groups and associated to adjacent chromosome blocks, is crucial in order to get a representative multi-locus marker genotype. In fact, microsatellite retrieved from non-coding regions (genomic SSR markers) meet this requirement more precisely than those derived from expressed regions (EST-SSR markers). Nevertheless, the application of EST-SSR markers cannot be excluded when phylogenetic relationships have to be investigated. It is well known that SSR markers belonging to coding regions may be functionally more informative than those deriving from non-coding ones, because associated with transcribed regions of the genome and thus reflecting the genetic diversity within genes or adjacent to genes [149]. Moreover, the association with trait loci with Mendelian inheritance is particularly requested in case of needs for marker-assisted selection (MAS).

About the localization of target microsatellites in the cellular genomes, nuclear SSR (nSSR) markers seem to be more polymorphic than plastidial and mitochondrial ones (cpSSR and mtSSR markers) and because of their co-dominance, the former are the only markers useful for assessing

the genetic value of breeding stocks, even if the abundance and the haploid nature of the latter ones make them particularly suitable for phylogenetic and genetic diversity studies.

As far as the microsatellite repeat is concerned, the most recommended motifs are di-nucleotide and tri-nucleotide repeats, whereas mononucleotide repeats need caution because of technical drawbacks which can be experienced in the allele discrimination. SSR markers with tetra-nucleotide or more repeats display a polymorphism inversely proportional to the complexity of the motif. The so-called perfect SSR markers are preferred because of their ease of scorability. It is also worth emphasizing that the choice of SSR markers is also dependent on the occurrence of null alleles for a given locus and the informativeness in terms of allele diversity indexes. First of all, any rate of null alleles can underestimate heterozygosity and affect the reliability of the analysis. Second, the calculation of some informative indexes cannot be underrated: it represents a crucial step of the planning of any analysis. What emerges from the 90 studies here reviewed is a lack of wider consensus among the authors regarding the best informative index to calculate and this makes the comparison difficult also among studies performed on the same cultivars and with the same markers. The power of discrimination (PD), the confusion probability (C), the polymorphism information content (PIC), the probability of identity (PI), the total probability of identity (PI_t) and the probability of null allele (r) are all parameters able to describe exhaustively the efficiency of the set of SSR markers used in a given species.

In conclusion, there is the urgent need to establish a common procedure for SSR genotyping with a universal set of marker loci to be analysed in each species. In parallel, the reference varieties must be defined in each species in order to maximize not only the reproducibility but also the portability of marker data, aware that the residual variation in laboratory procedures and equipment cannot be completely avoided.

References

1. Food and Agriculture Organization of the United Nations Statistics Division [Internet]. 2016 [cited 2016 May 30]. Available from: http://faostat3.fao.org/mes/methodology_list/E
2. Food and Agriculture Organization of the United Nations. Value of Agricultural Production [Internet]. 2016 [cited 2016 May 30]. Available from: <http://faostat3.fao.org/download/Q/QV/E>
3. Science IS for H. International Code of Nomenclature for Cultivated Plants. 9th ed. 2016. 190 p.
4. List of Italian PDO and PGI [Internet]. [cited 2016 May 30]. Available from: <https://www.politicheagricole.it/flex/cm/pages/ServeBLOB.php/L/IT/IDPagina/2090>
5. Sector plan, collaboration among MIPAAF, CRA, INEA and ISMEA [Internet]. [cited 2016 May 30]. Available from: <http://www.pianidisetto.it/flex/FixedPages/Common/RegistroVarietale.php/L/IT>
6. National register of horticultural varieties: Ministry of Agricultural, Food and Forestry Policies [Internet]. [cited 2016 May 30]. Available from: <http://www.politicheagricole.it/flex/cm/pages/ServeBLOB.php/L/IT/IDPagina/7051>
7. National register of vitis varieties. Ministry of Agricultural, Food and Forestry Policies [Internet]. [cited 2016 May 30]. Available from: <http://catalogoviti.politicheagricole.it/catalogo.php>
8. National register of agrarian varieties. Ministry of Agricultural, Food and Forestry Policies [Internet]. [cited 2016 May 30]. Available from: <http://www.politicheagricole.it/flex/cm/pages/ServeBLOB.php/L/IT/IDPagina/7051>
9. List of orientation among fruit trees: ongoing project between MIPAAF and regions [Internet]. [cited 2016 May 30]. Available from: <http://plantgest.imagelinenetwork.com>
10. List of italian DOC and DOCG wines, considered by the UE as PDO products [Internet]. [cited 2016 May 30]. Available from: <https://www.politicheagricole.it/flex/cm/pages/ServeBLOB.php/L/IT/IDPagina/4625>
11. Legislative Decree: Attuazione della direttiva 2008/90 relativa alla commercializzazione dei materiali di moltiplicazione delle piante da frutto destinate alla produzione di frutti. Article 7. Italy;
12. Lopez-Vizcón C, Ortega F. Detection of mislabelling in the fresh potato retail market employing microsatellite markers. *Food Control* [Internet]. 2012 Aug [cited 2014 Dec 12];26(2):575–9. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0956713512000941>
13. Mahajan V, Jakse J, Havey MJ, Lawande KE. Genetic fingerprinting of onion cultivars using SSR markers. *Indian J Hortic*. 2009;66(1):62–8.
14. Muñoz-Falcón JE, Vilanova S, Plazas M, Prohens J. Diversity, relationships, and genetic fingerprinting of the Listada de Gandía eggplant landrace using genomic SSRs and EST-SSRs. *Sci Hortic (Amsterdam)*. 2011;129(2):238–46.
15. De Mattia F, Imazio S, Grassi F, Lovicu G, Tardaguila J, Failla O, et al. Genetic characterization of sardinia grapevine cultivars by SSR markers analysis. *J Int des Sci la Vigne du Vin*. 2007;41(4):175–84.
16. Rao R, Bencivenni M, La Mura M, Araujo-Burgos T, Corrado G. Molecular characterisation of Vesuvian apricot cultivars: Implications for the certification and authentication of protected plant material. *J Hortic Sci Biotechnol*. 2010;85(1):42–7.
17. Motilal L, Butler D. Verification of identities in global cacao germplasm collections. *Genet Resour Crop Evol*. 2003;50(8):799–807.
18. Dossett M, Bassil N V, Finn CE. SSR Fingerprinting of Black Raspberry Cultivars Shows Discrepancies in Identification. 2012;49–54.
19. Benjak A, Ercisli S, Vokurka A, Maletić E, Pejić I. Genetic relationships among grapevine cultivars native to Croatia, Greece and Turkey. *Vitis - J Grapevine Res*. 2005;44(2):73–7.

20. Zhou L, Matsumoto T, Tan H-W, Meinhardt LW, Mischke S, Wang B, et al. Developing single nucleotide polymorphism markers for the identification of pineapple (*Ananas comosus*) germplasm. *Hortic Res* [Internet]. 2015;2(October):15056. Available from: <http://www.nature.com/articles/hortres201556>
21. Chandra A, Grisham MP, Pan Y. Allelic divergence and cultivar-specific SSR alleles revealed by capillary electrophoresis using fluorescence-labeled SSR markers in sugarcane. *Genome* [Internet]. 2014;57(6):363–72. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25247737>
22. Akkaya MS, Buyukunal-Bal EB. Assessment of genetic variation of bread wheat varieties using microsatellite markers. *Euphytica* [Internet]. 2004;135(2):179–85. Available from: <http://link.springer.com/10.1023/B:EUPH.0000014908.02499.41>
23. Rekik I, Salimonti A, Kamoun NG, Muzzalupo I, Lepais O, Gerber S, et al. Characterization and identification of tunisian olive tree varieties by microsatellite markers. *HortScience*. 2008;43(5):1371–6.
24. Martínez LE, Cavagnaro PF, Masuelli RW, Zúñiga M. SSR-based assessment of genetic diversity in South American *Vitis vinifera* varieties. *Plant Sci* [Internet]. 2006 Jun [cited 2014 Dec 12];170(6):1036–44. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0168945205004292>
25. Pasqualone A, Di Rienzo V, Nasti R, Blanco A, Gomes T, Montemurro C. Traceability of Italian Protected Designation of Origin (PDO) table olives by means of microsatellite molecular markers. *J Agric Food Chem* [Internet]. 2013 Mar 27;61(12):3068–73. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23461435>
26. Pasqualone A, Alba V, Mangini G, Blanco A, Montemurro C. Durum wheat cultivar traceability in PDO Altamura bread by analysis of DNA microsatellites. *Eur Food Res Technol*. 2010;230(5):723–9.
27. Caramante M, Corrado G, Monti LM, Rao R. Simple sequence repeats are able to trace tomato cultivars in tomato food chains. *Food Control* [Internet]. 2011 Mar [cited 2014 Dec 12];22(3–4):549–54. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0956713510003348>
28. Wadl PA, Wang X, Moulton JK, Hokanson SC, Skinner JA, Rinehart TA, et al. Transfer of *Cornus florida* and *C. kousa* Simple Sequence Repeats to Selected *Cornus* (Cornaceae) Species. *J Am Soc Hortic Sci*. 2010;135(3):279–88.
29. Laciš G, Rashaļ I, Rūisa S, Trajkovski V, Iezzoni AF. Assessment of genetic diversity of Latvian and Swedish sweet cherry (*Prunus avium* L.) genetic resources collections by using SSR (microsatellite) markers. *Sci Hortic (Amsterdam)*. 2009;121(4):451–7.
30. Korinsak S, Sriprakhon S, Sirithanya P, Jairin J, Vanavichit a., Toojinda T. Identification of microsatellite markers (SSR) linked to a new bacterial blight resistance gene xa33 (t) in rice cultivar “Ba7.” *Maejo Intern J Sci Technol*. 2009;3(2):235–47.
31. Fan S, Bielenberg DG, Zhebentyayeva TN, Reighard GL, Okie WR, Holland D, et al. Mapping quantitative trait loci associated with chilling requirement, heat requirement and bloom date in peach (*Prunus persica*). *New Phytol*. 2010;185(4):917–30.
32. Ashkani S, Rafii MY, Rusli I, Sariah M, Abdullah SNA, Rahim HA, et al. SSRs for Marker-Assisted Selection for Blast Resistance in Rice (*Oryza sativa* L.). *Plant Mol Biol Report*. 2012;30(1):79–86.
33. Cheta Kumar, MR; Vishwanath K; Shivakumar N; Rajendra Prasad S; Rahda BR. Utilization of SSR Markers for Seed Purity Testing in Popular Rice Hybrids (*Oryza sativa* L.). *Ann plant Sci*. 2012;1(1):1–5.
34. Ibañez J, Van Eeuwijk FA. Microsatellite profiles as a basis for intellectual property protection in grape. *Acta Hortic*. 2003;603:41–7.
35. Fan L, Zhang MY, Liu QZ, Li LT, Song Y, Wang LF, et al. Transferability of Newly Developed Pear SSR Markers to Other Rosaceae Species. *Plant Mol Biol Report*. 2013;31(6):1271–82.
36. Satya P, Paswan PK, Ghosh S, Majumdar S, Ali N. Confamilial transferability of simple sequence repeat (SSR) markers from cotton (*Gossypium hirsutum* L.) and jute (*Corchorus olitorius* L.) to

twenty two Malvaceous species. 3 Biotech [Internet]. 2016;6(1):65. Available from: <http://link.springer.com/10.1007/s13205-016-0392-z>

37. Emanuelli F, Lorenzi S, Grzeskowiak L, Catalano V, Stefanini M, Troglio M, et al. Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. *BMC Plant Biol* [Internet]. 2013;13(1):39. Available from: <http://www.biomedcentral.com/1471-2229/13/39>
38. Filippi C V, Aguirre N, Rivas JG, Zubrzycki J, Puebla A, Cordes D, et al. Population structure and genetic diversity characterization of a sunflower association mapping population using SSR and SNP markers. *BMC Plant Biol* [Internet]. 2015;15(1):52. Available from: <http://www.biomedcentral.com/1471-2229/15/52>
39. Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, et al. Current trends in microsatellite genotyping. *Mol Ecol Resour*. 2011;11(4):591–611.
40. Masi P, Spagnoletti Zeuli PL, Donini P. Development and analysis of multiplex microsatellite markers sets in common bean (*Phaseolus vulgaris* L.). *Mol Breed*. 2003;11(4):303–13.
41. Ganai MW, Röder MS. Microsatellite and SNP Markers in Wheat Breeding. *Genomics*. 2007;2:1–24.
42. Gonzaga ZJ. Evaluation of SSR and SNP Markers for Molecular Breeding in Rice. *Plant Breed Biotechnol* [Internet]. 2015;3(2):139–52. Available from: http://www.plantbreedbio.org/journal/view.html?uid=185&page=&sort=&scale=10&all_k=&s_t=&s_a=&s_k=&s_v=3&s_n=2&spage=&pn=search&year=&vmd=Full
43. Sánchez-Pérez R, Ruiz D, Dicenta F, Egea J, Martínez-Gómez P. Application of simple sequence repeat (SSR) markers in apricot breeding: Molecular characterization, protection, and genetic relationships. *Sci Hortic (Amsterdam)*. 2005;103(3):305–15.
44. Sardaro MLS, Marmioli M, Maestri E, Marmioli N. Genetic characterization of Italian tomato varieties and their traceability in tomato food products-Sardaro-2012-Food Science & Nutrition-Wiley Online Library. *Food Sci Nutr* [Internet]. 2013 Jan [cited 2014 Dec 1];1(1):54–62. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3951568&tool=pmcentrez&rendertype=abstract>
45. Baleiras-Couto MM, Eiras-Dias JE. Detection and identification of grape varieties in must and wine using nuclear and chloroplast microsatellite markers. *Anal Chim Acta* [Internet]. 2006 Mar [cited 2014 Nov 13];563(1–2):283–91. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0003267005016922>
46. Pasqualone A, Montemurro C, Caponio F, Blanco A. Identification of virgin olive oil from different cultivars by analysis of DNA microsatellites. *J Agric Food Chem* [Internet]. 2004 Mar 10;52(5):1068–71. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/14995099>
47. Laurens F, Durel CE, Lascostes M. Molecular characterization of French local apple cultivars using SSRs. *Acta Hortic*. 2004;663:639–42.
48. Ercisli S, Ipek A, Barut E. SSR marker-based DNA fingerprinting and cultivar identification of olives (*Olea europaea*). *Biochem Genet* [Internet]. 2011 Oct [cited 2014 Dec 12];49(9–10):555–61. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21476017>
49. Montemurro C, Simeone R, Pasqualone A, Ferrara E, Blanco A. Genetic relationships and cultivar identification among 112 olive accessions using AFLP and SSR markers. *J Hortic Sci Biotechnol*. 2005;80(1):105–10.
50. Díaz A, De la Rosa R, Martín A, Rallo P. Development, characterization and inheritance of new microsatellites in olive (*Olea europaea* L.) and evaluation of their usefulness in cultivar identification and genetic relationship studies. *Tree Genet Genomes* [Internet]. 2006 Mar 28 [cited 2014 Dec 12];2(3):165–75. Available from: <http://link.springer.com/10.1007/s11295-006-0041-5>
51. Taamalli W, Geuna F, Bassi D, Daoud D, Zarrouk M. SSR marker based DNA fingerprinting of

- Tunisian olive (*Olea europaea* L.) varieties. Vol. 7, Journal of Agronomy. 2008. p. 176–81.
52. Doveri S, Sabino Gil F, Díaz A, Reale S, Busconi M, da Câmara Machado A, et al. Standardization of a set of microsatellite markers for use in cultivar identification studies in olive (*Olea europaea* L.). *Sci Hortic (Amsterdam)* [Internet]. 2008 May [cited 2014 Dec 11];116(4):367–73. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0304423808000496>
 53. Baldoni L, Cultrera NG, Mariotti R, Ricciolini C, Arcioni S, Vendramin GG, et al. A consensus list of microsatellite markers for olive genotyping. *Mol Breed* [Internet]. 2009 May 13 [cited 2014 Dec 12];24(3):213–31. Available from: <http://link.springer.com/10.1007/s11032-009-9285-8>
 54. Muzzalupo I, Stefanizzi F, Perri E. Evaluation of olives cultivated in southern Italy by simple sequence repeat markers. *HortScience*. 2009;44(3):582–8.
 55. Muzzalupo I, Stefanizzi F, Salimonti A, Falabella R, Perri E. Microsatellite Markers for Identification of a Group of Italian Olive Accessions. *Sci Agric*. 2009;66(5):685–90.
 56. Bracci T, Sebastiani L, Busconi M, Fogher C, Belaj A, Trujillo I. SSR markers reveal the uniqueness of olive cultivars from the Italian region of Liguria. *Sci Hortic (Amsterdam)* [Internet]. 2009 Sep [cited 2014 Dec 9];122(2):209–15. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0304423809002155>
 57. Alba V, Montemurro C, Sabetta W, Pasqualone A, Blanco A. SSR-based identification key of cultivars of *Olea europaea* L. diffused in Southern-Italy. *Sci Hortic (Amsterdam)* [Internet]. 2009 Dec [cited 2014 Dec 12];123(1):11–6. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0304423809003550>
 58. Corrado G, Imperato A, la Mura M, Perri E, Rao R. Genetic diversity among olive varieties of southern Italy and the traceability of olive oil using SSR markers. *J Hortic Sci Biotechnol*. 2011;86(5):461–6.
 59. Ipek A, Barut E, Gulen H, Ipek M. Assessment of inter- and intra-cultivar variations in olive using SSR markers. *Sci Agric [Internet]*. 2012;69(5):327–35. Available from: <http://www.scopus.com/inward/record.url?eid=2-s2.0-84867483027&partnerID=40&md5=c6c81882560aca1bd542c465b04eb2f1>
 60. Muzzalupo I, Salimonti A, Stefanizzi F, Falabella R, Perri E. Microsatellite Markers for Characterization and Identification of Olive (*Olea europaea*) Cultivars in South Italy. *Vi Int Symp Olive Grow* [Internet]. 2012;949:67–70. Available from: https://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=5&SID=V1H3TbxBXvPOXeKSSLj&page=1&doc=1
 61. Las Casas G, Scollo F, Distefano G, Continella A, Gentile A, La Malfa S. Molecular characterization of olive (*Olea europaea* L.) Sicilian cultivars using SSR markers. *Biochem Syst Ecol* [Internet]. 2014 Dec [cited 2014 Dec 4];57:15–9. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0305197814002026>
 62. Abdessemed S, Muzzalupo I, Benbouza H. Assessment of genetic diversity among Algerian olive (*Olea europaea* L.) cultivars using SSR marker. *Sci Hortic (Amsterdam)* [Internet]. 2015;192:10–20. Available from: <http://dx.doi.org/10.1016/j.scienta.2015.05.015>
 63. Sakar E, Unver H, Ercisli S. Genetic Diversity Among Historical Olive (*Olea europaea* L.) Genotypes from Southern Anatolia Based on SSR Markers. *Biochem Genet* [Internet]. 2016;54(6):842–53. Available from: <http://link.springer.com/10.1007/s10528-016-9761-x>
 64. Caramante M, Rao R, Monti LM, Corrado G. Discrimination of “San Marzano” accessions: A comparison of minisatellite, CAPS and SSR markers in relation to morphological traits. *Sci Hortic (Amsterdam)* [Internet]. 2009 May [cited 2014 Dec 12];120(4):560–4. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0304423808005347>
 65. Meng F, Xu X, Huang F, Li J. Analysis of Genetic Diversity in Cultivated and Wild Tomato Varieties in Chinese Market by RAPD and SSR. *Agric Sci China* [Internet]. 2010 Oct [cited 2014 Dec 12];9(10):1430–7. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1671292709602340>

66. El-awady MAM, El-tarras AAE, Hassan MM. Genetic diversity and DNA fingerprint study in tomato (*Solanum lycopersicum* L.) cultivars grown in Egypt using simple sequence repeats (SSR) markers. *African J Biotechnol.* 2012;11(96):16233–40.
67. Todorovska E, Ivanova A, Ganeva D, Pevicharova G, Molle E, Bojinov B, et al. Assessment of genetic variation in Bulgarian tomato (*Solanum lycopersicum* L.) genotypes, using fluorescent SSR genotyping platform. *Biotechnol Biotechnol Equip* [Internet]. 2014 Jun 4 [cited 2014 Dec 12];28(1):68–76. Available from: <http://www.tandfonline.com/doi/abs/10.1080/13102818.2014.901683>
68. Korir NK, Diao W, Tao R, Li X, Kayesh E, Li a, et al. Genetic diversity and relationships among different tomato varieties revealed by EST-SSR markers. *Genet Mol Res* [Internet]. 2014 Jan;13(1):43–53. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24446286>
69. Miskoska-Milevska E, Popovski Z, Dimitrievska B, Bandzo K. DNA microsatellite analysis for tomato genetic differentiation. *Genetika* [Internet]. 2015;47(3):1123–30. Available from: <http://www.doiserbia.nb.rs/Article.aspx?ID=0534-00121503123M>
70. Mercati F, Longo C, Poma D, Araniti F, Lupini A, Mammano MM, et al. Genetic variation of an Italian long shelf-life tomato (*Solanum lycopersicon* L.) collection by using SSR and morphological fruit traits. *Genet Resour Crop Evol* [Internet]. 2014;62(5):721–32. Available from: <http://link.springer.com/10.1007/s10722-014-0191-5>
71. Zulini L, Russo M, Peterlunger E. Genotyping wine and table grape cultivars from Apulia (southern Italy) using microsatellite markers. *Vitis.* 2002;41(4):183–7.
72. This P, Jung a, Boccacci P, Borrego J, Botta R, Costantini L, et al. Development of a standard set of microsatellite reference alleles for identification of grape cultivars. *Theor Appl Genet* [Internet]. 2004 Nov [cited 2014 Dec 9];109(7):1448–58. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15565426>
73. Hvarleva T, Hadjinicoli A, Atanassov I, Atanassov A, Ioannou N. Genotyping *Vitis vinifera* L. cultivars of Cyprus by microsatellite analysis. *Vitis.* 2005;44(2):93–7.
74. Cipriani G, Marrazzo MT, Di Gaspero G, Pfeiffer A, Morgante M, Testolin R. A set of microsatellite markers with long core repeat optimized for grape (*Vitis* spp.) genotyping. *BMC Plant Biol* [Internet]. 2008 Jan [cited 2014 Dec 12];8:127. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2625351&tool=pmcentrez&rendertype=abstract>
75. Salmaso M, Forestan C, Varotto S, Lucchin M. Biodiversity of local Italian grapevine cultivars detected by SSR markers. *ISHS Acta Hort: Proc. IXth Intl. Conf. on Grape Genetics and Breeding.* 2009;827:137–42.
76. Marinoni DT, Raimondi S, Ruffa P, Lacombe T, Schneider A. Identification of grape cultivars from Liguria (north-western Italy). *Vitis - J Grapevine Res.* 2009;48(4):175–83.
77. Jahnke G, Májer J, Lakatos a., Molnár JG, Deák E, Stefanovits-Bányai É, et al. Isoenzyme and microsatellite analysis of *Vitis vinifera* L. varieties from the Hungarian grape germplasm. *Sci Hortic (Amsterdam)* [Internet]. 2009 Apr [cited 2014 Dec 12];120(2):213–21. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0304423808004470>
78. Zoghalmi N, Riahi L, Laucou V, Lacombe T, Mliki a., Ghorbel a., et al. Origin and genetic diversity of Tunisian grapes as revealed by microsatellite markers. *Sci Hortic (Amsterdam)* [Internet]. 2009 May [cited 2014 Dec 12];120(4):479–86. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S030442380800527X>
79. Veloso MM, Almandanim MC, Baleiras-Couto M, Pereira HS, Carneiro LC, Feveireiro P, et al. Microsatellite database of grapevine (*Vitis vinifera* L.) cultivars used for wine production in Portugal. *Cienc e Tec Vitivinic.* 2010;25(2):53–61.
80. Cipriani G, Marrazzo MT, Peterlunger E. Molecular characterization of the autochthonous grape cultivars of the region Friuli Venezia Giulia - North-Eastern Italy. *Vitis - J Grapevine Res.*

2010;49(1):29–38.

81. Cipriani G, Spadotto A, Jurman I, Di Gaspero G, Crespan M, Meneghetti S, et al. The SSR-based molecular profile of 1005 grapevine (*Vitis vinifera* L.) accessions uncovers new synonymy and parentages, and reveals a large admixture amongst varieties of different geographic origin. *Theor Appl Genet* [Internet]. 2010 Nov [cited 2014 Dec 12];121(8):1569–85. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20689905>
82. Jahnke G, Májer J, Varga P, Szőke B. Analysis of clones of Pinots grown in Hungary by SSR markers. *Sci Hortic (Amsterdam)* [Internet]. 2011 May [cited 2014 Dec 12];129(1):32–7. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0304423811001129>
83. Moreno-Sanz P, Loureiro MD, Suárez B. Microsatellite characterization of grapevine (*Vitis vinifera* L.) genetic diversity in Asturias (Northern Spain). *Sci Hortic (Amsterdam)* [Internet]. 2011 Jun [cited 2014 Dec 12];129(3):433–40. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0304423811002019>
84. Díaz-Losada E, Tato Salgado a., Ramos-Cabrer a. M, Díaz-Hernández B, Pereira-Lorenzo S. Genetic and geographical structure in grapevines from northwestern Spain. *Ann Appl Biol* [Internet]. 2012 Jul 10 [cited 2014 Dec 12];161(1):24–35. Available from: <http://doi.wiley.com/10.1111/j.1744-7348.2012.00548.x>
85. Guo DL, Zhang Q, Zhang GH. Characterization of grape cultivars from China using microsatellite markers. *Czech J Genet Plant Breed.* 2013;49(4):164–70.
86. Basheer-Salimia R, Lorenzi S, Batarseh F, Moreno-Sanz P, Emanuelli F, Grando MS. Molecular identification and genetic relationships of Palestinian grapevine cultivars. *Mol Biotechnol* [Internet]. 2014 Jun [cited 2014 Dec 12];56(6):546–56. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24469973>
87. Galbacs Z, Molnar S, Halasz G, Kozma P, Hoffmann S, Kovacs L, et al. Identification of grapevine cultivars using microsatellite-based DNA barcodes. *Vitis.* 2009;48(1):17–24.
88. Zarouri B, Vargas AM, Gaforio L, Aller M, de Andrés MT, Cabezas JA. Whole-genome genotyping of grape using a panel of microsatellite multiplex PCRs. *Tree Genet Genomes.* 2015;11:17.
89. Lei W, Juan Z, Linde L, Li Z, Lijuan W, Dechang H. Genetic diversity of grape germplasm as revealed by microsatellite (SSR) markers. *African J Biotechnol* [Internet]. 2015;14(12):990–8. Available from: <http://academicjournals.org/journal/AJB/article-abstract/DA5410651677>
90. Dograr N, Akin-Yalin S AM. Discriminating durum wheat cultivars using highly polymorphic simple sequence repeat DNA markers. *Plant Breed.* 1999;119:884–6.
91. Prasad M, Varshney RK, Roy JK, Balyan HS, Gupta PK. The use of microsatellites for detecting DNA polymorphism, genotype identification and genetic diversity in wheat. *TAG Theor Appl Genet* [Internet]. 2000 Feb;100(3–4):584–92. Available from: <http://link.springer.com/10.1007/s001220050077>
92. Röder MS, Wendehake K, Korzun V, Bredemeijer G, Laborie D, Bertrand L, et al. Construction and analysis of a microsatellite-based database of European wheat varieties. *Theor Appl Genet* [Internet]. 2002 Dec [cited 2014 Dec 12];106(1):67–73. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12582872>
93. Leigh F, Lea V, Law J, Wolters P, Powell W, Donini P. Assessment of EST- and genomic microsatellite markers for variety discrimination and genetic diversity studies in wheat. *Euphytica.* 2003;133(3):359–66.
94. Perry DJ. Identification of Canadian durum wheat varieties using a single PCR. *Theor Appl Genet.* 2004;109(1):55–61.
95. Fujita Y, Fukuoka H, Yano H. Identification of wheat cultivars using EST–SSR markers. *Breed Sci* [Internet]. 2009;59(2):159–67. Available from: <http://joi.jlc.jst.go.jp/JST.JSTAGE/jsbbs/59.159?from=CrossRef>

96. Schuster I, Vieira ESN, da Silva GJ, Franco FDA, Marchioro VS. Genetic variability in Brazilian wheat cultivars assessed by microsatellite markers. *Genet Mol Biol.* 2009;32(3):557–63.
97. Salem KFM, Röder MS, Börner A. Assessing genetic diversity of Egyptian hexaploid wheat (*Triticum aestivum* L.) using microsatellite markers. *Genet Resour Crop Evol.* 2015;62(3):377–85.
98. Henkrar F, El-Haddoury J, Ouabbou H, Nsarellah N, Iraqi D, Bendaou N, et al. Genetic diversity reduction in improved durum wheat cultivars of Morocco as revealed by microsatellite markers. *Sci Agric [Internet]*. 2016;73(2):134–41. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-90162016000200134&lng=en&nrm=iso&tlng=en
99. Goulão L, Oliveira C. Molecular characterisation of cultivars of apple (*Malus × domestica* Borkh.) using microsatellite (SSR and ISSR) markers. *Euphytica [Internet]*. 2001;122:81–9. Available from: <http://link.springer.com/article/10.1023/A:1012691814643>
100. Kitahara K, Matsumoto S, Yamamoto T, Soejima J, Abe K. Molecular Characterization of Apple Cultivars in Japan by. *Sci Technol.* 2005;130(6):885–92.
101. Galli Z, Halász G, Kiss E, Heszky L, Dobránszki J. Molecular identification of commercial apple cultivars with microsatellite markers. *HortScience.* 2005;40(7):1974–7.
102. Melchiade D, Foroni I, Corrado G, Santangelo I, Rao R. Authentication of the “Annurca” Apple in Agro-food Chain by Amplification of Microsatellite Loci. *Food Biotechnol [Internet]*. 2007 Mar 6 [cited 2014 Dec 12];21(1):33–43. Available from: <http://www.tandfonline.com/doi/abs/10.1080/08905430701191114>
103. Garkava-Gustavsson L, Kolodinska Brantestam A, Sehic J, Nybom H. Molecular characterisation of indigenous Swedish apple cultivars based on SSR and S-allele analysis. *Hereditas.* 2008;145(3):99–112.
104. Pereira-Lorenzo S, Ramos-Cabrera M, González-Díaz J, Díaz-Hernández MB. Genetic assessment of local apple cultivars from La Palma, Spain, using simple sequence repeats (SSRs). *Sci Hortic (Amsterdam) [Internet]*. 2008 Jun [cited 2014 Dec 12];117(2):160–6. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0304423808001039>
105. Patocchi a., Fernández-Fernández F, Evans K, Gobbin D, Rezzonico F, Boudichevskaia a., et al. Development and test of 21 multiplex PCRs composed of SSRs spanning most of the apple genome. *Tree Genet Genomes [Internet]*. 2008 Aug 29 [cited 2014 Dec 12];5(1):211–23. Available from: <http://link.springer.com/10.1007/s11295-008-0176-7>
106. van Treuren R, Kemp H, Ernsting G, Jongejans B, Houtman H, Visser L. Microsatellite genotyping of apple (*Malus × domestica* Borkh.) genetic resources in the Netherlands: application in collection management and variety identification. *Genet Resour Crop Evol [Internet]*. 2010 Jan 28 [cited 2014 Nov 30];57(6):853–65. Available from: <http://link.springer.com/10.1007/s10722-009-9525-0>
107. Baric S, Wagner J, Storti A, Via JD. Application of an Extended Set of Microsatellite DNA Markers for the Analysis of Presumed Synonym Cultivars of Apple. *Acta Hortic.* 2011;918:303–8.
108. Moriya S, Iwanami H, Okada K, Yamamoto T, Abe K. A practical method for apple cultivar identification and parent-offspring analysis using simple sequence repeat markers. *Euphytica [Internet]*. 2010 Nov 5 [cited 2014 Dec 12];177(1):135–50. Available from: <http://link.springer.com/10.1007/s10681-010-0295-8>
109. Liu GS, Zhang YG, Tao R, Fang JG, Dai HY. Identification of apple cultivars on the basis of simple sequence repeat markers. *Genet Mol Res [Internet]*. 2014 Jan;13(3):7377–87. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25222236>
110. Omasheva ME, Chekalin S V., Galiakparov NN. Evaluation of molecular genetic diversity of wild apple *Malus sieversii* populations from Zailiysky Alatau by microsatellite markers. *Russ J Genet [Internet]*. 2015;51(7):647–52. Available from: <http://link.springer.com/10.1134/S1022795415070108>

111. Lassois L, Denancé C, Ravon E, Guyader A, Guisnel R, Hibrand-Saint-Oyant L, et al. Genetic Diversity, Population Structure, Parentage Analysis, and Construction of Core Collections in the French Apple Germplasm Based on SSR Markers. *Plant Mol Biol Report* [Internet]. 2016;34(4):827–44. Available from: <http://dx.doi.org/10.1007/s11105-015-0966-7>
112. Bigliuzzi J, Scali M, Paolucci E, Cresti M, Vignani R. DNA Extracted with Optimized Protocols Can Be Genotyped to Reconstruct the Varietal Composition of Monovarietal Wines. *Am J Enol Vitic* [Internet]. 2012 Dec 1 [cited 2014 Dec 4];63(4):568–73. Available from: <http://www.ajevonline.org/cgi/doi/10.5344/ajev.2012.12014>
113. Boccacci P, Akkak A, Torello Marinoni D, Gerbi V, Schneider A. Genetic traceability of Asti Spumante and Moscato d’Asti musts and wines using nuclear and chloroplast microsatellite markers. *Eur Food Res Technol* [Internet]. 2012 Jul 7 [cited 2014 Dec 12];235(3):439–46. Available from: <http://link.springer.com/10.1007/s00217-012-1770-3>
114. Pereira L, Martins-Lopes P, Batista C, Zanol GC, Clímaco P, Brazão J, et al. Molecular Markers for Assessing Must Varietal Origin. *Food Anal Methods* [Internet]. 2012 Feb 11 [cited 2014 Nov 30];5(6):1252–9. Available from: <http://link.springer.com/10.1007/s12161-012-9369-7>
115. Recupero M, Garino C, De Paolis A, Cereti E, Coisson J-D, Travaglia F, et al. A Method to Check and Discover Adulteration of Nebbiolo-Based Monovarietal Musts: Detection of Barbera and Dolcetto cv via SSR Analysis Coupled with Lab-On-Chip® Microcapillary Electrophoresis. *Food Anal Methods* [Internet]. 2012 Sep 16 [cited 2014 Dec 12];6(3):952–62. Available from: <http://link.springer.com/10.1007/s12161-012-9506-3>
116. Muccillo L, Gambuti A, Frusciante L, Iorizzo M, Moio L, Raieta K, et al. Biochemical features of native red wines and genetic diversity of the corresponding grape varieties from Campania region. *Food Chem* [Internet]. 2014 Jan 15 [cited 2014 Dec 12];143:506–13. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24054274>
117. Doveri S, O’Sullivan DM, Lee D. Non-concordance between genetic profiles of olive oil and fruit: A cautionary note to the use of DNA markers for provenance testing. *J Agric Food Chem*. 2006;54(24):9221–6.
118. Pasqualone A, Montemurro C, Summo C, Sabetta W, Caponio F, Blanco A. Effectiveness of microsatellite DNA markers in checking the identity of protected designation of origin extra virgin olive oil. In: *Journal of Agricultural and Food Chemistry*. 2007;55(10):3857–62.
119. Martins-Lopes P, Gomes S, Santos E, Guedes-Pinto H. DNA markers for Portuguese olive oil fingerprinting. *J Agric Food Chem*. 2008;56(24):11786–91.
120. Alba V, Sabetta W, Blanco A, Pasqualone A, Montemurro C. Microsatellite markers to identify specific alleles in DNA extracted from monovarietal virgin olive oils. *Eur Food Res Technol* [Internet]. 2009 Apr 11 [cited 2014 Dec 12];229(3):375–82. Available from: <http://link.springer.com/10.1007/s00217-009-1062-8>
121. Zohreh Rabiei, Sattar Tahmasebi Enferadi, Abbas Saidi, Sonia Patui, GianPaolo Vannozzi. Simple sequence repeats amplification: a tool to survey the genetic background of olive oils. *Iran J Biotechnol*. 2010;8:24–31.
122. Vietina M, Agrimonti C, Marmiroli M, Bonas U, Marmiroli N. Applicability of SSR markers to the traceability of monovarietal olive oils. *J Sci Food Agric* [Internet]. 2011 Jun [cited 2014 Dec 12];91(8):1381–91. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21384371>
123. Rotondi A, Beghè D, Fabbri A, Ganino T. Olive oil traceability by means of chemical and sensory analyses: A comparison with SSR biomolecular profiles. *Food Chem* [Internet]. 2011 Dec [cited 2014 Dec 11];129(4):1825–31. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0308814611008259>
124. Ben-Ayed R, Grati-Kamoun N, Sans-Grout C, Moreau F, Rebai A. Characterization and authenticity of virgin olive oil (*Olea europaea* L.) cultivars by microsatellite markers. *Eur Food Res Technol* [Internet]. 2011 Nov 30 [cited 2014 Dec 12];234(2):263–71. Available from:

<http://link.springer.com/10.1007/s00217-011-1631-5>

125. Pérez-Jiménez M, Besnard G, Dorado G, Hernandez P. Varietal tracing of virgin olive oils based on plastid DNA variation profiling. *PLoS One* [Internet]. 2013 Jan [cited 2014 Dec 12];8(8):e70507. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3737381&tool=pmcentrez&rendertype=abstract>
126. Falagas ME, Pitsouni EI, Malietzis G a, Pappas G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB J* [Internet]. 2008;22(2):338–42. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17884971>
127. Khan MA, Han Y, Zhao YF, Troggio M, Korban SS. A multi-population consensus genetic map reveals inconsistent marker order among maps likely attributed to structural variations in the apple genome. *PLoS One* [Internet]. 2012 Jan [cited 2014 Dec 12];7(11):e47864. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3489900&tool=pmcentrez&rendertype=abstract>
128. Genome database for Rosaceae [Internet]. [cited 2016 May 30]. Available from: <http://www.rosaceae.org/search/markers>
129. Jaillon O, Aury J-. M, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* [Internet]. 2007;449. Available from: <http://dx.doi.org/10.1038/nature06148>
130. Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, et al. A physical map of the 1Gb bread wheat chromosome 3B. *Science* (80-). 2008;322(2008):101–4.
131. Muleo R, Morgante M, Velasco R, Cavallini A, Perrotta G BL. Olive Tree Genomic. In: *Olive Germplasm - The Olive cultivation, table olive and olive oil industry in Italy*. 2012. Rijeka, Croatia: InTech.
132. Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, et al. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* [Internet]. 2012;485(7400):635–41. Available from: <http://dx.doi.org/10.1038/nature11119>
http://www.nature.com/nature/journal/v485/n7400/full/nature11119.html?WT.ec_id=NATURE-20120531
133. Cruz F, Julca I, Gómez-Garrido J, Loska D, Marcet-Houben M, Cano E, et al. Genome sequence of the olive tree, *Olea europaea*. *Gigascience* [Internet]. 2016;5(1):29. Available from: <http://gigascience.biomedcentral.com/articles/10.1186/s13742-016-0134-5>
134. *Olea* database [Internet]. [cited 2016 May 30]. Available from: http://www.oleadb.it/ssr/ssr_dca_search.php
135. Tomato microsatellite database [Internet]. [cited 2016 May 30]. Available from: <http://webapp.cabgrid.res.in/tomsatdb/tutorial.html>
136. Tomato genomic resources database [Internet]. [cited 2016 May 30]. Available from: <http://59.163.192.91/tomato2/ssr.html>
137. Tomato - Kazusa Marker Database [Internet]. [cited 2016 May 30]. Available from: <http://marker.kazusa.or.jp/tomato>
138. Grape Microsatellite Collection [Internet]. [cited 2016 May 30]. Available from: <http://meteo.iasma.it/genetica/gmc.html>
139. This P, Dettweiler E. EU-Project Genres CT96 No81 : European Vitis Database and Results Regarding the Use of a Common Set of Microsatellite Markers. 2003.
140. Italian Vitis Database [Internet]. [cited 2016 May 30]. Available from: <http://www.vitisdb.it/descriptors/microsatellites>
141. European Vitis Database [Internet]. [cited 2016 May 30]. Available from: <http://www.eu->

vitis.de/index.php

142. Bennett MD, Leitch IJ. Nuclear-DNA amounts in angiosperms. *Ann bot.* 1995;76:113–176.
143. Wheat Microsatellite Consortium [Internet]. [cited 2016 May 30]. Available from: <http://wheat.pw.usda.gov/ggpages/SSR/WMC>
144. Li GY, Dreisigacker S, Warburton ML, Xia XC, He ZH SQ. Development of a fingerprinting database and assembling an SSR reference kit for genetic diversity analysis of wheat. *Acta Agron Sin.* 2006;32(12):1771–8.
145. Han Y, Korban SS. An overview of the apple genome through BAC end sequence analysis. *Plant Mol Biol.* 2008;67(6):581–8.
146. HiDRAS SSR database [Internet]. [cited 2016 May 30]. Available from: <http://www.hidras.unimi.it/HiDRAS-SSRdb/pages/index.php>
147. Lateur M, Ordidge M, Engels J, Lipman E. Report of a Working Group on Malus / Pyrus. Report of a Working Group on Malus/Pyrus. Fourth Meeting, 7-9 March 2012, Weggis, Switzerland. 2012.
148. Tanaka TS, Kobayashi FU, Joshi GIRIPR, Onuki RI, Sakai HI, Nasuda SH, et al. Next-Generation Survey Sequencing and the Molecular Organization of Wheat Chromosome 6B }. *DNA Res.* 2014;(October 2013):103–14.
149. Varshney RK, Graner A, Sorrells ME. Genic microsatellite markers in plants: Features and applications. *Trends Biotechnol.* 2005;23(1):48–55.
150. Hu J, Wang L, Li J. Comparison of genomic SSR and EST-SSR markers for estimating genetic diversity in cucumber. *Biol Plant.* 2011;55(3):577–80.
151. Nunzia Scotti TC and LM. Mitochondrial DNA and RNA Isolation from Small Amounts of Potato Tissue. *Plant Mol Biol Report.* 2001;19(March):67.
152. Wolfe KH, Li W-H, Sharp PM. Rates of Nucleotide Substitution Vary Greatly among Plant Mitochondrial, Chloroplast, and Nuclear DNAs. *Proc Natl Acad Sci U S A* [Internet]. 1987;84(24):9054–8. Available from: <http://www.jstor.org/stable/30764>
153. Borgo R, Souty-Grosset C, Bouchon D, Gomot L. PCR-RFLP Analysis of Mitochondrial DNA for Identification of Snail Meat Species. *J Food Sci.* 1996;61(1):1–4.
154. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* [Internet]. 1980;32(3):314–31. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1686077&tool=pmcentrez&rendertype=abstract>
155. Nagy S, Poczai P, Cernák I, Gorji AM, Hegedus G, Taller J. PICcalc: An online program to calculate polymorphic information content for molecular genetic studies. *Biochem Genet.* 2012;50(9–10):670–2.
156. Anderson JA, Churchill GA, Autrique JE, Tanksley SD, Sorrells ME. Optimizing parental selection for genetic linkage maps. *Genome Dyn.* 1993;36(i):181–6.
157. Aranzana MJ, Carbó J, Arús P. Microsatellite variability in peach [*Prunus persica* (L.) Batsch]: cultivar identification, marker mutation, pedigree inferences and population structure. *Theor Appl Genet* [Internet]. 2003 May [cited 2014 Dec 12];106(8):1341–52. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12750778>
158. Qanbari S, Eskandari Nasab MP, Osfoori R HNA. Power of Microsatellite Markers for Analysis of Genetic Variation and Parentage Verification in Sheep. *Pak J Biol Sci.* 2007;10:1632–8.
159. Silfverberg-Dilworth E, Matasci CL, Van De Weg WE, Van Kaauwen MPW, Walser M, Kodde LP, et al. Microsatellite markers spanning the apple (*Malus x domestica* Borkh.) genome. *Tree Genet Genomes.* 2006;2(4):202–24.
160. Urquhart A, Kimpton CP, Downes TJ, Gill P. Variation in short tandem repeat sequences--a survey of

- twelve microsatellite loci for use as forensic identification markers. *Int J Legal Med.* 1994;107:13–20.
161. Brookfield JF. A simple new method for estimating null allele frequency from heterozygote deficiency. *Mol Ecol.* 1996;5(3):453–5.
 162. Sehic J, Garkava-Gustavsson L, Nybom H. More harmonization needed for DNA-based identification of apple germplasm. *Acta Hortic.* 2013;976:277–84.
 163. Sefc KM, Lopes MS, Mendonça D, Rodrigues Dos Santos M, Laimer Da Camara Machado M DCMA. Identification of microsatellite loci in olive (*Olea europaea* L.) and their characterization in Italian and Iberian olive trees. *Mol Ecol* [Internet]. 2000;9:1433–49. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/10972783>
 164. Carriero F, Fontanazza G, Cellini F, Giorio G. Identification of simple sequence repeats (SSRs) in olive (*Olea europaea* L.). *Theor Appl Genet.* 2002;104(2–3):301–7.
 165. Cipriani G, Marrazzo MT, Marconi R, Cimato A, Testolin R. Microsatellite markers isolated in olive (*Olea europaea* L.) are suitable for individual fingerprinting and reveal polymorphism within ancient cultivars. *Theor Appl Genet.* 2002;104(2–3):223–8.
 166. de la Rosa R, James CM TK. Isolation and characterization of polymorphic microsatellites in olive (*Olea europaea* L.) and their transferability to other genera in the Oleaceae. *Mol Ecol Notes.* 2002;2:265–7.
 167. Olive-Track. Traceability of origin and authenticity of olive oil by combined genomic and metabolomic approaches [Internet]. [cited 2016 May 30]. Available from: <http://www.dsa.unipr.it/foodhealth/oliv-track/index.html>
 168. Marmioli N, Maestri E, Pafundo S, Vietina M, Biotechnology E. Molecular traceability of olive oil: From plant genomics to food genomics. In: *Advances in Olive Resources.* 2009. Kerala, India: Transworld Research Network.
 169. Agrimonti C, Vietina M, Pafundo S, Marmioli N. The use of food genomics to ensure the traceability of olive oil. *Trends Food Sci Technol* [Internet]. 2011 May [cited 2014 Dec 11];22(5):237–44. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S092422441100029X>
 170. Caroli S, Santoni S, Ronfort J. AMaCAID: A useful tool for Accurate Marker Choice for Accession Identification and Discrimination. *Mol Ecol Resour.* 2011;11(4):733–8.
 171. Arias RS, Ballard LL, Scheffler BE. UPIC: Perl scripts to determine the number of SSR markers to run. *Bioinformatics* [Internet]. 2009;3(8):352–60. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19707300>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2720665>

Chapter II

Developing a molecular identification assay of old landraces for the genetic authentication of typical agro-food products: the case study of the barley ‘Agordino’

Abstract

The orzo Agordino is a very old local variety of domesticated barley (*Hordeum vulgare* ssp. *distichum* L.) that is native to the Agordo District, Province of Belluno, and is widespread in the Veneto Region, Italy. Seeds of this landrace are widely used for the preparation of very famous dishes of the dolomitic culinary tradition such as barley soups, bakery products and local beer. Understanding the genetic diversity and identity of the Agordino barley landrace is a key step to establish conservation and valorisation strategies of this local variety and also to provide molecular traceability tools useful to ascertain the authenticity of its derivatives. The gene pool of the Agordino barley landrace was reconstructed using 60 phenotypically representative individual plants and its genotypic relationships with commercial varieties were investigated using 21 pure lines widely cultivated in the Veneto Region. For genomic DNA analysis, following an initial screening of 14 mapped microsatellite (SSR) loci, seven discriminant markers were selected on the basis of their genomic position across linkage groups and polymorphic marker alleles per locus. The genetic identity of the local barley landrace was determined by analysing all SSR markers in a single multi-locus PCR assay. Extent of genotypic variation within the Agordino barley landrace and the genotypic differentiation between the landrace individuals and the commercial varieties was determined. Then, as few as four highly informative SSR loci were selected and used to develop a molecular traceability system exploitable to verify the genetic authenticity of food products deriving from the Agordino landrace. This genetic authentication assay was validated using both DNA pools from individual Agordino barley plants and DNA samples from Agordino barley food products. On the whole, our data support the usefulness and robustness of this DNA-based diagnostic tool for the orzo Agordino identification, which could be rapidly and efficiently exploited to guarantee the authenticity of local varieties and the typicality of food products.

Keywords: microsatellites, genotyping, landraces, traceability, barley, food authentication

Introduction

A local variety or landrace is a dynamic and ancient population of a cultivated crop, characterized by a well established historical and geographical identity, which is locally adapted to the natural resources, agronomic practices of farmers and cultural traditions of consumers [1]. Within the last few decades, interest in local varieties has been renewed primarily because of the rediscovery of traditional local food products and their potential economic value on different market scales.

Among cultivated species, the landraces of barley (*Hordeum vulgare* L.), a herbaceous plant from the Poaceae family that is widely distributed and primarily used for feed, food and malt production, offer a notable case study. Barley is the fourth most important cereal crop in the world, following wheat (*Triticum* spp.), rice (*Oryza* spp.) and corn (*Zea mays* L.). It is cultivated primarily in the temperate regions of Asia, Europe and North America, with a total area of 56 million hectares worldwide and an annual production of 144 million tonnes [2].

The barley landrace Agordino is a two-row local variety of *Hordeum vulgare* that is widespread in the Veneto Region (Italy) and native to the Agordino District (Belluno). Across the entire dolomitic region, as reported in a nineteenth century manuscript [3], the cultivation of this landrace is a centuries-old tradition. Currently, this landrace survives only in small plots of land totalling a few hectares 1500 m above sea level, particularly in the municipality of Livinallongo del Col di Lana (the province of Belluno) and in some fields of the Belluno and Feltre Valleys. Morphologically, this barley landrace has very tall stems, normally exceeding 100 cm in height, and has modest yields (approx. 2.6 t/ha) when compared with modern two-row varieties [4]. The seeds of this landrace are widely used for the preparation of the barley soup ('zuppa d'orzo'), probably the most famous local dish of the dolomitic culinary tradition. Flour obtained from the stone milling of Agordino barley is used for bakery products, such as barley bread and cookies, and roasted seeds are particularly appreciated to produce barley coffee ('caffè d'orzo'). Moreover, some breweries recently started to commercialize lines of beer dedicated to this local variety.

In recent decades, molecular markers have been successfully used not only to characterize and preserve commercial and local barley varieties but also to authenticate raw materials and food derivatives [5]. Among the PCR-based techniques, microsatellite or simple sequence repeat (SSR) markers provide a fully codominant and highly polymorphic marker system, widely exploited to implement reproducible and transferable molecular assays, which have found utility in detecting inter- and intra-population differences [6].

In this study, the genetic identity of the Agordino barley landrace has been assessed with SSR markers. As preliminary analysis, we assembled molecular data in order to determine the population genetic structure of the Agordino barley landrace and its genetic relationships with several commercial varieties. Then, we exploited this crucial information to develop a genetic traceability assay of the food products deriving from the Agordino barley landrace. In fact, the most informative SSR marker loci allowed us to determine the most common, highly shared or fixed, marker alleles and varietal genotypes, and to validate marker allele combinations (*i.e.* multi-locus haplotypes), which are typical and specifically associated with the Agordino barley landrace and, thus, to its food derivatives. The description of the genetic traits of this barley landrace, historically cultivated in the Belluno provincial area, north-eastern Italy, supports a larger adoption of molecular marker analyses for the identification of local materials and the authentication of their seed lots and food derivatives.

Materials and Methods

Plant material and genomic DNA isolation

A total of 60 individual samples of the Italian barley landrace Agordino (OA) were collected from experimental populations originally collected at different locations across the Belluno province and maintained at the experimental station of the Institute N. Strampelli (Lonigo, VI, Italy), as part of a BIO.NET research project funded by the Rural Development Program of the Veneto Region. Moreover, 35 individual samples belonging to 21 pure lines of the commercial varieties most

commonly used by farmers in Veneto were also obtained from the Institute (Table 1), and used as reference standards and test samples. Because the commercial varieties were represented by pure lines and therefore were composed of populations of plants sharing the same genotype, one or two individuals only were analysed for each variety.

Table 1. List of barley materials characterised using molecular markers, including 60 individuals of the Agordino barley landrace and 21 commercial varieties chosen among the most commonly used by farmers in Veneto, Italy, for a total of 95 samples

Variety	N (Sample)	Variety	N (Sample)
Agordino	60	Calanque	1
Concerto	2	Casanova	2
Scarlett	2	Cometa	2
Arda	2	Flanelle	2
Baraka	2	Kangoo	1
Barberusse	2	Marjorie	1
Plaisant	2	Sfera	1
Leonessa	2	Alba	1
Saxsonia	2	Scandella	2
Atomo	2	Tunika	1
Braemar	2	Primus	1

Genomic DNA was isolated from 100 mg of fresh leaf tissue using a DNeasy plant kit (Qiagen, Valencia, CA, USA), following the procedure provided by the suppliers. The integrity of extracted DNA samples was estimated by electrophoresis on a 0.8 % agarose/1× TAE gel containing 1× SYBR Safe DNA stain (Life Technologies, Carlsbad, CA, USA). Both the purity and quantity of DNA extracts were assessed with a NanoDrop 2000c UV-Vis spectrophotometer (Thermo Scientific, Pittsburgh, PA, USA).

Analysis of SSR markers

To genotype the 95 barley samples, PCR amplifications and microsatellite (SSR) marker analyses were performed using the M13-tailed SSR method described by Schuelke [7], with some modifications. An M13-labelled primer (5'-TTGTAACGACGGCCAGT-3') was used in combination with a specific SSR-targeting forward primer with a 5'-M13 tail and a specific SSR-targeting reverse primer.

The set of 14 SSR marker loci investigated in this study was obtained from [8]. An initial screening of a subset of 12 DNA samples, randomly chosen among the local and commercial varieties, was performed to investigate the amplification efficiency and polymorphism information content of the chosen SSR markers (Table 2). One SSR marker locus per linkage group was then selected for the genotyping of all DNA samples, including Bmag0872, Bmag0125, EBmac0871, Bmag0808, EBmatc0003, Bmac0727 and Bmag0321 [8].

Table 2. Information on the simple sequence repeat (SSR) marker loci analysed in this study, including linkage group, locus name, primers sequence and amplicon size. The polymorphism information content (PIC) value is also reported for each locus [8]. Loci written in bold were the ones assayed in this study

Linkage group	Locus	Forward	Reverse	Size bp	PIC
1H	Bmag0872	ATGTACCATTACGCATCCA	GAAATGTAGAGATGGCACTTG	125-153	0.81
1H	Bmag0211	GCAAGCTTCCTAAATCCTTA	TGCAGACAGTTTTTCATATACA	174-220	0.83
2H	Bmag0125	AATTAGCGAGAACAAAATCAC	AGATAACGATGCACCACC	138-157	0.76
2H	EBmac0415	GAAACCCATCATAGCAGC	AAACAGCAGCAAGAGGAG	247-282	0.58
3H	EBmac0871	TGCCCTGTGTGTTATTGT	CCCCAAGTGAACATTGAC	194-211	0.83
3H	Bmac0127b	AACTATGTCCAGTCGTTTCC	CTTGTCGATCATCTTATTCAGA	118	0.67
4H	Bmag0808	TCATAGACTACGACGAAGATG	TCTTTGGATGTGTGTTACTG	199-209	0.84
4H	Bmag0490	TGATACATCAAGATCGTGACA	GGGACTGAGTGTATGAATGAG	121	0.73
5H	EBmatc0003	AATTTTGCAAAGCTGGAGG	CATTATGGTGGGGTTCATGT	113-129	0.60
5H	EBmac0824	ATTCATCGATCTTGTATTAGTCC	ACATCATGTCGATCAAAGC	308-371	0.43
6H	Bmac0727	AACTATGTCCAGTCGTTTCC	CTTGTCGATCATCTTATTCAGA	126-140	0.83
6H	Bmag0613	AAGAACCCATATGATCCAAC	CTCCATGACTATGAGGAGAAG	171-218	0.73
7H	Bmag0321	ATTATCTCCTGCAACAACCTA	CTCCGGAACACTACGACAAG	230-243	0.70
7H	Bmag0206	TTTTCCCCTATTATAGTGACG	TAGAACTGGGTATTTTCCTGA	239	0.79

The PCR reaction consisted of a 20- μ L final volume that contained 1x Platinum[®] Multiplex PCR Master Mix, 10 % GC Enhancer (Applied Biosystems, Carlsbad, CA, USA), 0.25 μ M of each tailed primer, 0.75 μ M of each reverse primer, 0.5 μ M of each labelled primer (Applied Biosystems), 10 ng of DNA and distilled water. Amplifications were performed using a 9600 thermal cycler (Applied Biosystems) with a 96-well plate under the following conditions: 2 min at 95 °C, followed by 5 cycles at 95 °C for 30 s and at 60 °C for 90 s, which decreased by 0.8 °C with each cycle, and at 72 °C for 45 s, then 30 cycles at 95 °C for 30 s, at 56 °C for 90 s, and at 72 °C for 45 s. The reaction was terminated with a final extension of 10 min at 72 °C. The PCR products were then subjected to capillary electrophoresis with an ABI PRISM 3130xl Genetic Analyzer (Applied

Biosystems). The LIZ500 was adopted as molecular mass standard. Finally, the size of each peak was determined using Peak Scanner™ software v. 1.0 [9].

Marker analysis

To estimate the marker allele variation of the selected SSR loci in the 95 barley accessions, the polymorphism information content (PIC) was calculated using PICcalc software [10]. This index provides an estimate of the discriminating power of each co-dominant marker locus and it depends on the number of detectable marker alleles and on the distribution of their frequency.

Marker allele frequencies computed for the Agordino population were used to estimate the genetic diversity and differentiation statistics using the POPGENE software v. 1.32 [11]. The average number of alleles observed per locus (N_a) and the effective number of alleles per locus (N_e) were calculated according to Kimura and Crow [12]. For each marker locus and for all loci, the Nei's genetic diversity [13] of the Agordino landrace, which corresponds to the Nei's expected heterozygosity (H), was computed as:

$$H = 1 - \sum p_i^2 \quad (1)$$

being p_i the population frequency of the i^{th} marker allele. The polymorphism degree was calculated using Shannon's information index (I) of phenotypic diversity as follows [14]:

$$I = - \sum p_i^2 \ln p_i^2 \quad (2)$$

being p_i the population frequency of the i^{th} marker allele.

Genetic similarity estimates were also calculated between individuals of the local and the commercial accessions in all possible pairwise comparisons by applying the coefficient of simple matching. Differently from other statistics used to assess the genetic similarity between individuals, the simple matching coefficient, also known as Rohlf's coefficient [15], takes into account not only shared and polymorphic marker alleles, but also marker alleles missing in both samples under comparison and present in other samples of the same population, as factors that contribute to the

similarity estimates among pairwise combinations of individuals. The ordination analysis was performed according to the unweighted pair-group method with arithmetic mean (UPGMA), and the dendrogram and centroids of all accessions were constructed from the mean genetic similarity matrix. Principal coordinate analysis (PCoA) was applied to compute the first two principal components of the qualitative data matrix. All calculations and analyses were conducted using NTSYS-pc v. 2.21q [15]. A bootstrap statistical analysis was conducted to measure the stability of the computed branches with 1000 resampling replicates.

The genetic structure of the Agordino landrace and of the entire barley core collection based on a total of 95 DNA samples, including the individuals of the local population and commercial materials (*i.e.* pure lines), was also modelled using a Bayesian clustering algorithm implemented in STRUCTURE v. 2.2 [16]. Since barley (*Hordeum vulgare* L.) plants strictly reproduce by self-pollination and cultivated populations are composed of highly homozygous individuals, a haploid setting was used for this analysis. Using the admixture model with independent allele frequencies, ten replicate simulations were conducted for each value of number of populations (K), with K ranging from 1 to 20, using a burn-in of $2 \cdot 10^5$ and final run of 10^6 Markov Chain Monte Carlo (MCMC) steps [17]. The method described by Evanno *et al.* [18] was used to evaluate the most probable estimation of K.

Developmental validation tests

Genotypic data from all Agordino individuals showing a membership higher than 75 %, as assessed by STRUCTURE v. 2.2 [16], were used to identify the population-specific marker alleles at each locus and the frequency of the multi-locus haplotypes of the landrace. SSR loci showing at least one typical marker allele were selected for the validation procedure. Marker alleles were scored as ‘typical’ when present and shared by 100 % of the Agordino individuals and absent from the commercial varieties or with a frequency lower than 5 %.

For the validation study, three blends were prepared by combining genomic DNA of a randomly chosen Agordino genotype (sample OA23) with genomic DNA of a given commercial variety (namely ‘Arda’) in three different ratios: 1:1, 1:2 and 1:10. All DNA samples were then screened at the selected SSR marker loci in order to verify the detection of Agordino and non-Agordino marker alleles in experimental pooled conditions. In addition, genomic DNA extracted using cetyl trimethylammonium bromide CTAB method [19] from two commercial products of the Agordino barley, including barley seeds for soup preparation (Cooperativa Agricola La Fiorita, Cesiomaggiore BL, Italy) and barley crackers (Cooperativa Agricola La Fiorita), were analysed at the same SSR marker loci.

Results

Descriptive statistics of SSR marker loci

Levels of genetic variability among landrace individuals and commercial pure lines were high (Table 3). The polymorphism information content (PIC) value of the seven SSR marker loci chosen for the genotyping analysis was on average equal to 0.62, ranging from 0.50 (EBmatc0003) to 0.71 (Bmag0125). All the loci examined were polymorphic across all barley accessions, and the most common marker allele had a frequency of 0.60 (Ebmacc0871).

Table 3. Descriptive statistics of genetic diversity calculated across markers and barley accessions. Included are the frequency of the most common marker allele (p_i), average number of observed alleles (N_a) and of effective number of alleles (N_e) per locus, level of observed homozygosity (H_o), and estimates of Shannon’s information index of phenotypic diversity (I) and unbiased Nei’s genetic diversity (H). The polymorphism information content (PIC) coefficient was calculated for each locus. The overall values and standard deviations (S.D.) are also reported for each parameter

Locus ID	p_i	N_a	N_e	H_o	I	H	PIC
Bmag0872	0.45	7.00	3.54	1.00	1.49	0.72	0.68
Bmag0321	0.55	6.00	2.55	1.00	1.20	0.61	0.55
Bmag0808	0.33	5.00	3.65	1.00	1.38	0.73	0.68
Bmac0727	0.43	7.00	3.88	1.00	1.60	0.74	0.71
EBmac0871	0.60	6.00	2.42	1.00	1.16	0.59	0.54
EBmatc0003	0.59	3.00	2.30	1.00	0.96	0.56	0.50
Bmag0125	0.34	7.00	4.03	1.00	1.53	0.75	0.71
Agordino	0.93	3.86	2.13	1.00	0.84	0.45	
Commercial	0.82	4.43	2.78	1.00	1.13	0.60	
mean value	0.53	5.86	3.20	1.00	1.33	0.67	0.62
S.D.	0.19	1.46	0.75	0.00	0.23	0.08	0.09

Descriptive statistics of all SSR loci, in addition to the level of genetic diversity found across molecular markers and plant accessions, are reported in Table 3. In the barley populations, a total of 41 marker alleles were detected, and the average number of observed alleles (N_a) per SSR locus was equal to 5.86 for all accessions, with numbers that ranged from three to seven. In further detail, the average number of alleles per SSR locus was 3.86 for the Agordino landrace, whereas the mean was 4.43 for the set of commercial varieties. Moreover, the effective number of alleles (N_e) per SSR locus was 2.13 and 2.78 for the landrace and for the pure lines overall, respectively.

Estimates of both unbiased Nei's genetic diversity (H) and Shannon's information index of phenotypic diversity (I) were used to characterize the gene pools of the Agordino landrace and the commercial varieties (Table 3). The mean genetic diversity for all marker loci was 0.67 (S.D.=0.08) and ranged from the minimum of 0.56 (locus EBmatc0003) to the maximum of 0.75 (locus Bmag0125). The measures of molecular genetic diversity were 0.45 and 0.60 for the Agordino landrace and the commercial varieties, respectively. Shannon's information index (I) for all marker loci was 1.33 (S.D.=0.23) and ranged from the minimum of 0.96 (locus EBmatc0003) to the maximum of 1.60 (locus Bmac0727). This information index of marker phenotypic diversity was higher in the commercial accessions (1.13) than in the Agordino accessions (0.84; see Table 3). Finally, the Agordino landrace included a mixture of pure lines that were 100 % homozygous but for different marker alleles at one or more genomic loci. Each of the commercial varieties was shown to correspond to a single pure line and all these genetically different lines scored a homozygosity level of 100 %, as expected (Table 3).

Genetic diversity and cluster analysis

Genetic variability within and between varieties was investigated primarily by calculating genetic similarity estimates for all possible pairwise comparisons among the 95 DNA samples/genotypes using the entire set of marker alleles scored at all genomic loci. In particular, a pairwise genetic similarity matrix was calculated using a simple matching coefficient. Rohlf's genetic similarity

ranged from 61 to 100 % in the landrace population, with an average of 84 %, and the identical index varied from 60 to 100 %, with an average of 78 %, for the commercial varieties as a group. Between commercial varieties and the individuals of the Agordino landrace, genetic similarity values ranged between 50 and 90 %, with an average of 69 %.

Both ordination methods (UPGMA tree and PCoA centroids) based on the genetic similarity estimates identified two subgroups of individuals (Figures 1 and 2). Moreover, all 95 genotypes were divided into two major clusters, supported by a bootstrap value of 100 %, and most Agordino accessions were clustered into one well-defined major group, as shown in Figure 1. Most notably, the Agordino individuals contained at least 16 genetically different pure lines that were split into four subclusters, with bootstrap values ranging from 90 % to 100 %. Some of these pure lines were apparently overrepresented and comprised up to a dozen individuals each. The second major cluster included all commercial pure lines and four Agordino individuals (*i.e.* OA30, OA46, OA50 and OA52).

Principal coordinate analysis (PCoA) defined the centroids for all barley accessions, with the two-dimensional plot shown in Figure 2. Based on the marker alleles, the first principal coordinate accounted for 77.4% of the total variation and clearly separated the local genotypes from the commercial ones. In addition, consistently with the UPGMA tree, with PCoA four distinct subclusters were formed for the Agordino landrace population, with few individuals whose centroids were closely related to those of the commercial pure lines. The genotype of the commercial variety Alba proved to be very similar to that of some individuals of the local variety Agordino (see Figures 1 and 2). This finding was further supported by a mean genetic similarity value of 0.81 from the pair-wise comparison between the Alba accession and the Agordino population as a whole.

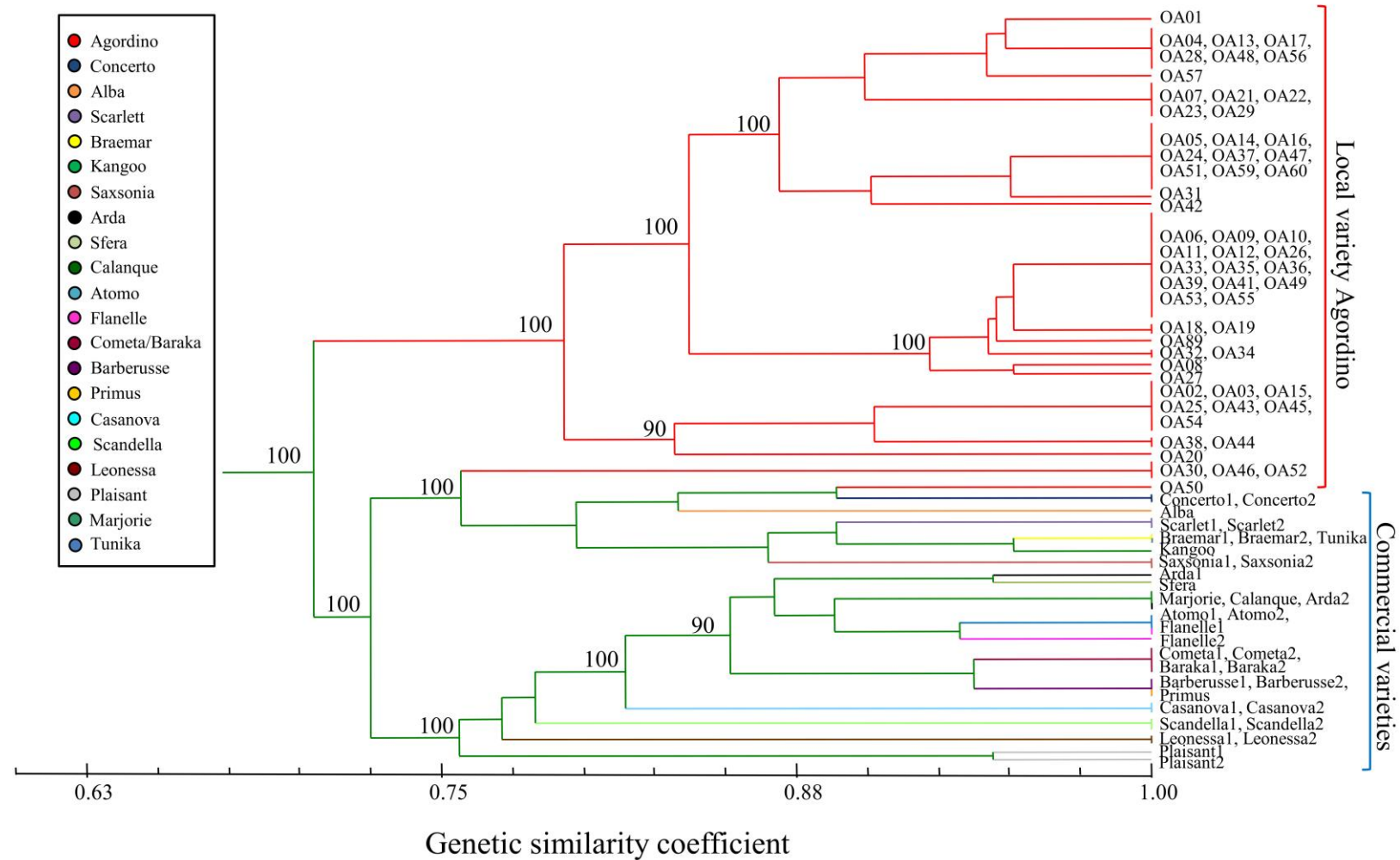


Figure 1. Unweighted pair-group method with arithmetic average mean UPGMA tree of the genetic similarity estimates computed among pairwise comparisons of barley accessions using the whole simple sequence repeat (SSR) marker data set, with nodes of the main subgroups supported by bootstrap values

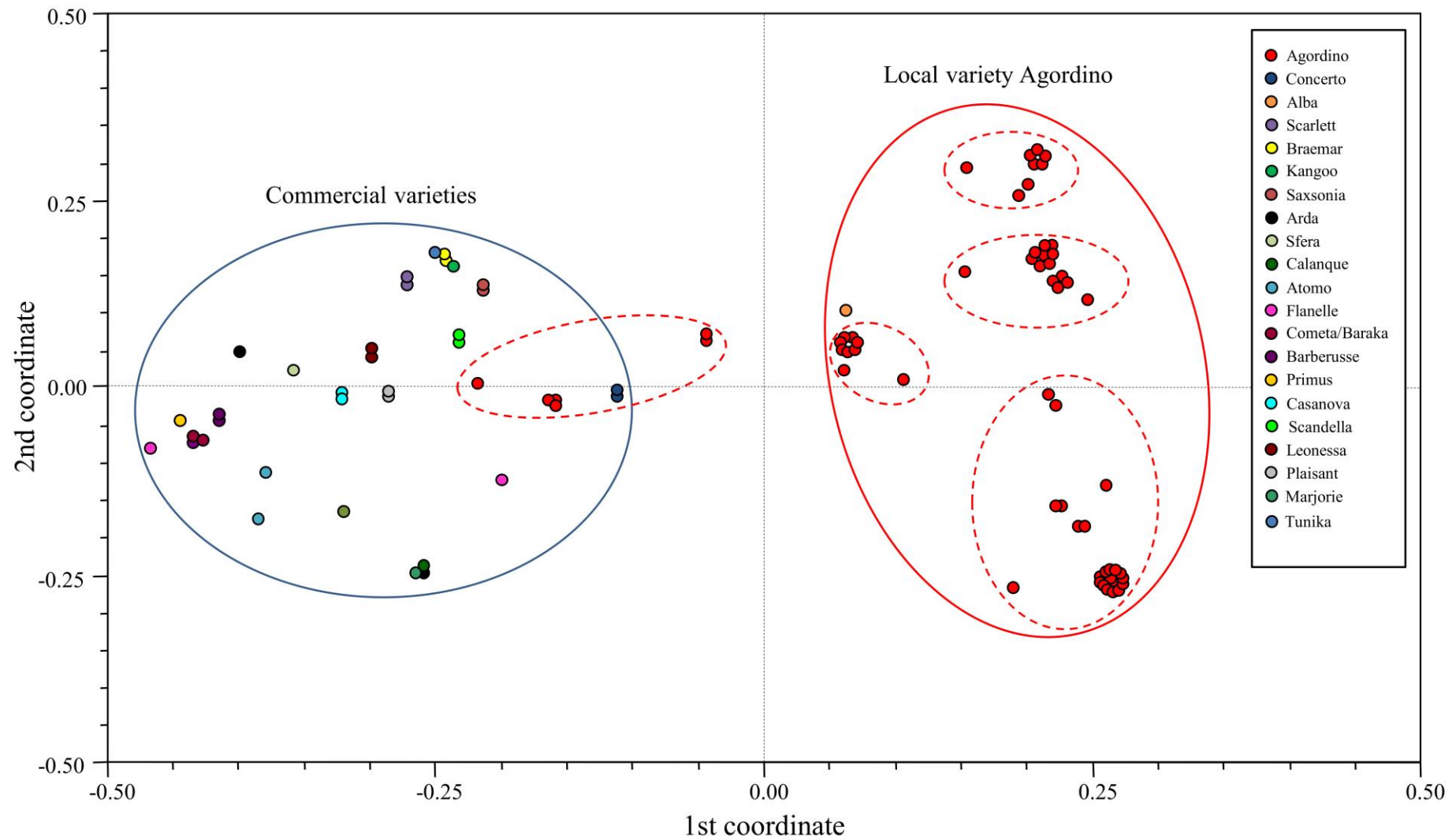


Figure 2. Two-dimensional centroids derived from the genetic similarity estimates computed among barley accessions in all possible pairwise comparisons using the whole simple sequence repeat (SSR) marker data set. Two main subgroups are distinguishable, one including most of the Agordino individuals, with a few exceptions, and the other containing all but one (Alba) commercial lines

Genetic structure analysis

Based on the marker alleles at all SSR loci, the genetic structure of the barley core collection was investigated using STRUCTURE v. 2.2. Following the procedure of Evanno *et al.*, examination of the population structure by estimation of ΔK values was consistent with the partitioning of the population into two genetically distinguishable subgroups (Figure 3), confirming the results from UPGMA analyses. In particular, the 95 barley samples were partitioned into two major marker allele clusters or ancestral multilocus haplotypes. For each plant accession, a vertical histogram was partitioned into $K=2$ coloured segments to represent the estimated membership of each hypothesized ancestral genotype. Single plant accessions were sorted by population type, and the clustering of individuals revealed only a few admixed accessions, such as OA20, OA38, OA44 (local variety) and Alba as commercial variety (see Figure 3), with membership values that ranged from 43 to 64 %. As the most dominant feature of the output, almost all individuals of the local landrace shared the same marker allele cluster, with accession scores of individual membership that were almost always higher than 94 %. By contrast, the apparently heterogeneous group of 21 commercial varieties was dominated by a single genetic pattern, providing strong support for these pure lines as a separate population with a distinct genetic background sharing a unique ancestry. Finally, four local samples (*i.e.*, OA30, OA46, OA50 and OA52) grouped closely with the cluster of commercial varieties.

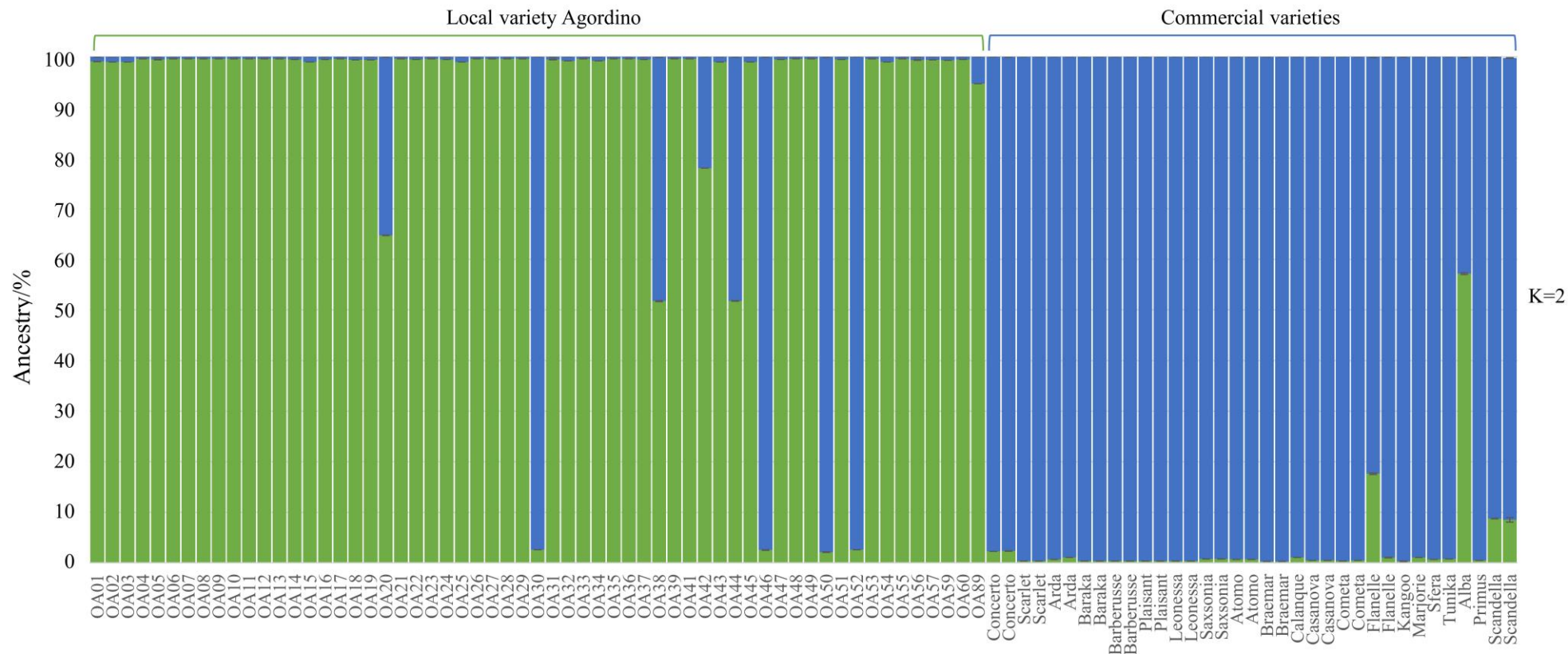


Figure 3. Population genetic structure of the Agordino landrace individuals ($N=60$) and the commercial varieties ($N=35$) as estimated by STRUCTURE v. 2.2 [16] using whole simple sequence repeat (SSR) marker data set. Each sample is represented by a vertical histogram partitioned into $K=2$ coloured segments that represent the estimated membership. The proportion of ancestry (%) is reported on the ordinate axis (the identification number of each accession is reported below each histogram)

Genetic authentication assay

In order to implement a genetic authentication assay, based on the genetic structure analysis, seven samples scoring a membership lower than 75 % to the founding group of the Agordino landrace (namely OA20, OA30, OA38, OA44, OA46, OA50 and OA52) were removed from further analysis because they were attributable to deviant or admixed haplotypes. The marker allele composition and proportion at each locus for the local population of Agordino barley is reported in Figure 4. Eight Agordino-specific marker alleles, defined as typical of this landrace because never detected in the commercial lines (Figure 4), were identified at four SSR loci, namely Bmag0872, Bmag0808, EBmatc0003 and Bmag0125.

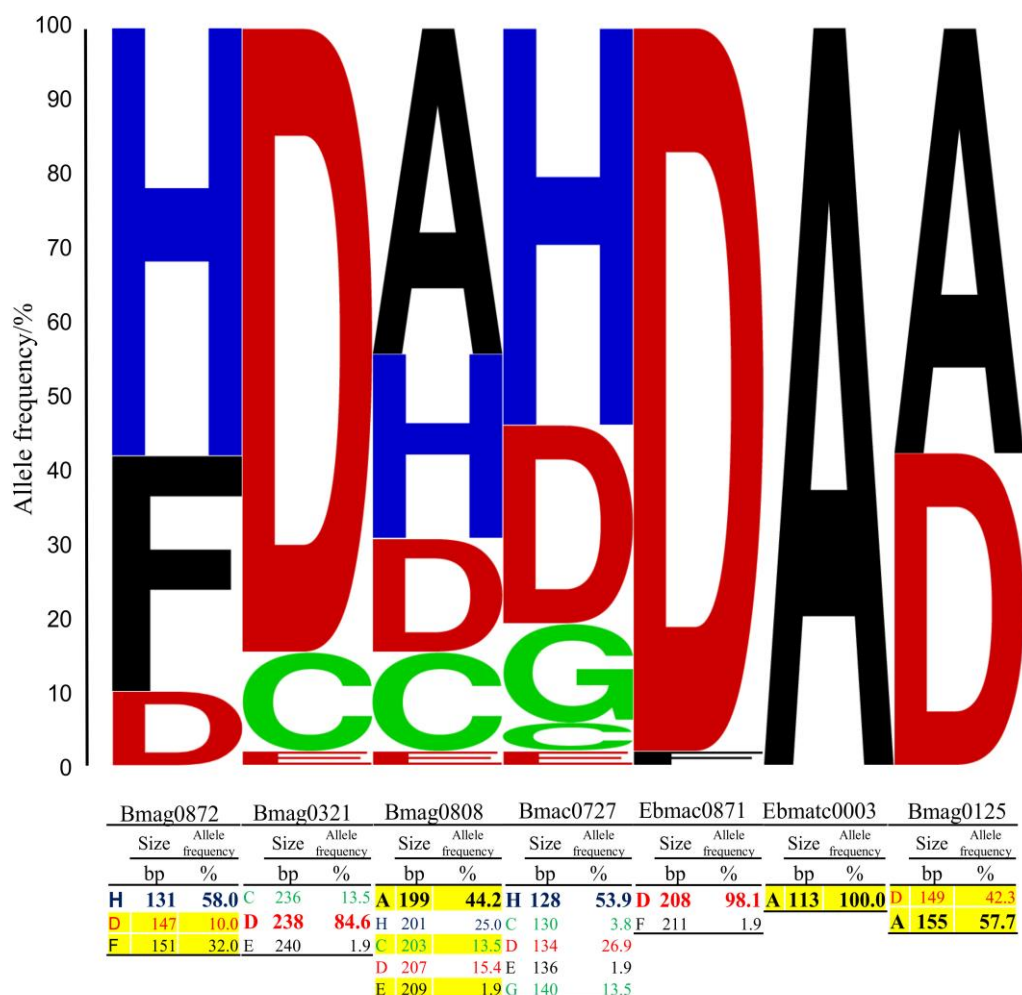


Figure 4. Graphical representation of the marker allele type and frequency at each SSR locus for the Agordino landrace. Each stack represents a locus and the height of letters within each stack indicates the relative frequency of the corresponding alleles. The correspondence between letters, allele sizes (bp) and frequencies (%) at each locus is reported in the table under the chart. For each locus, the most common allele is written in bold, whereas the typical alleles of the landrace are highlighted in yellow

An explanatory example for two of these loci is shown in Figure 5. Using the three blends of genomic DNA samples, the molecular phenotype (*i.e.* marker genotype) attributed to the Agordino germplasm (sample OA23, marker alleles of 199 and 155 bp) is clearly distinguishable from that of the commercial variety (line Arda, marker alleles of 201 and 157 bp), even when the concentration of the local variety is tenfold lower than the concentration of the commercial variety (see Figures 5c and f).

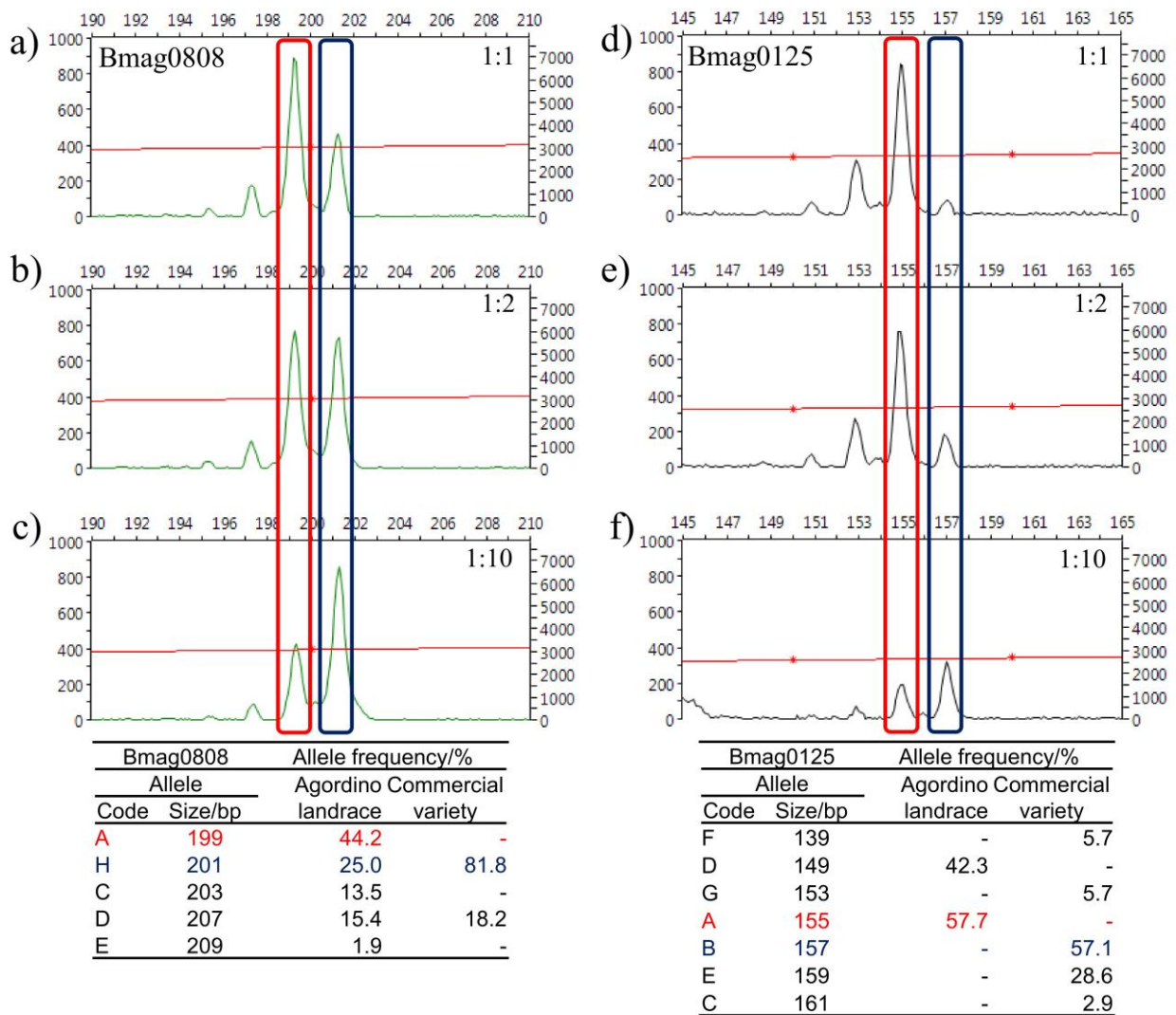


Figure 5. Electropherograms of microsatellite amplicons at loci Bmag0808 and Bmag0125 related to an experimental blend prepared by combining genomic DNA from an Agordino individual and the commercial variety Arda in three different ratios: 1:1 (a and d), 1:2 (b and e) and 1:10 (c and f). Rectangles highlight the marker alleles belonging to Agordino (red) and to Arda (blue). Tables report the correspondence among letters, allele sizes (bp) and allele frequencies (%) both in the Agordino population and the commercial group. Plots of the dye signal traces were obtained by Peak Scanner™ Software v. 1.0 (9)

Developmental validation tests were also performed successfully using commercial Agordino-based food products. Marker allelic profiles at the four informative loci using genomic DNA extracted from two commercial food products (*i.e.* barley seeds for soup preparation and barley crackers, respectively) as template for SSR amplifications are reported in Figure 6. For each locus, several marker alleles were detected but only some of them could be unambiguously assigned to the Agordino gene pool. It is worth mentioning that some of the marker alleles could not be univocally associated with the local variety because they were identified also in some of the commercial lines (Figure 6).

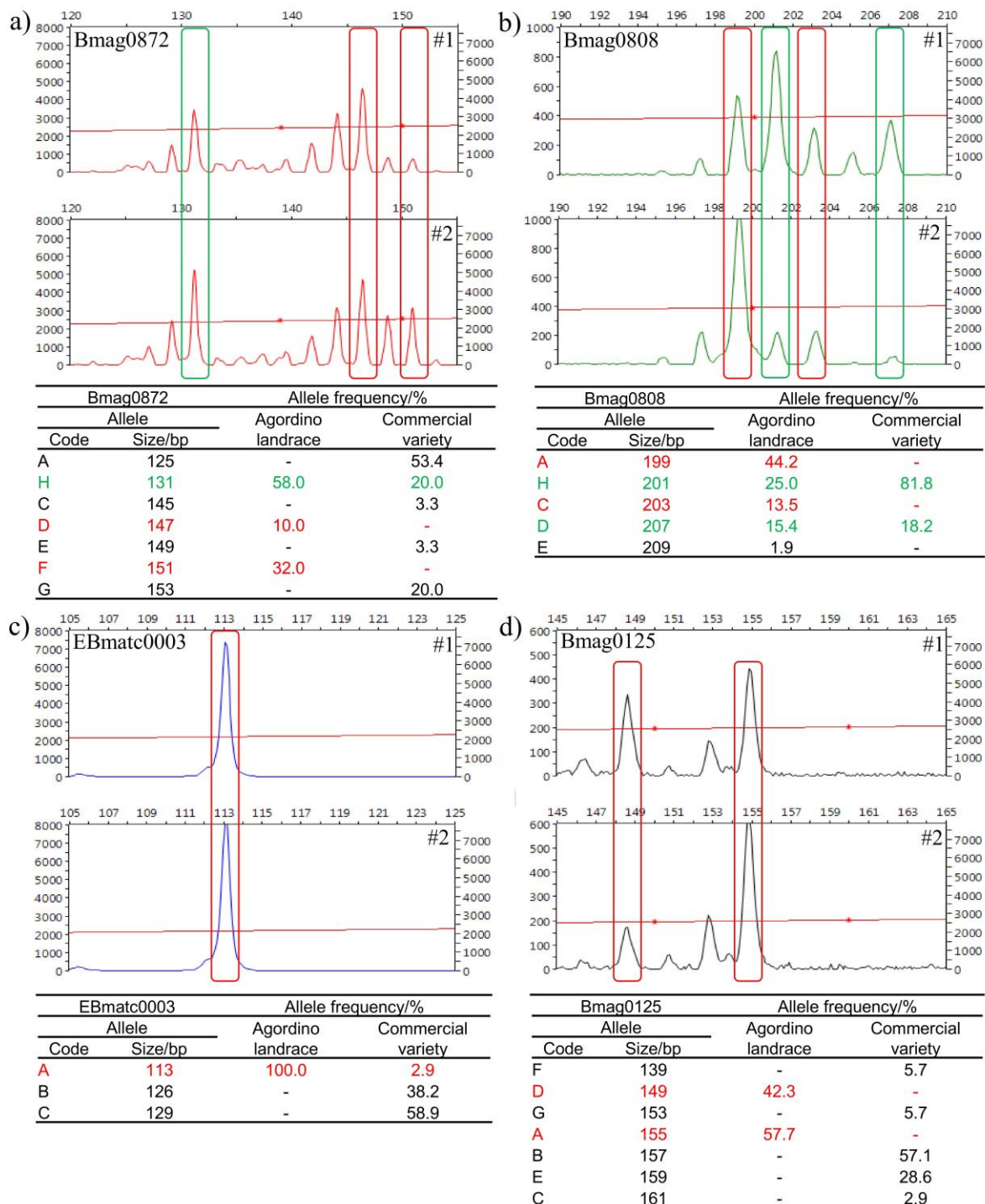


Figure 6. Electropherograms of microsatellite amplicons at four loci of genomic DNA extracted from two commercial food products deriving from Agordino, i.e. barley seeds for soup preparation (sample #1) and barley crackers (sample #2). Rectangles highlight the marker alleles unambiguously associated with the Agordino landrace, defined as typical (red), and the marker alleles shared with commercial varieties (green). Tables report the correspondence among letters, allele sizes (bp) and allele frequencies (%) both in the Agordino population and the commercial group. Plots of the dye signal traces were obtained by Peak Scanner™ Software v. 1.0 [9]

Discussion

The genetic identity and population structure of the Agordino landrace was assessed as preliminary goal to acquire essential information for the development of a molecular traceability assay of food products. This research reports original findings from the first analysis of the genetic diversity and identity of the old Italian landrace of barley locally named Agordino. On the basis of these findings, our goal was that of selecting a set of markers useful for the identification of plant materials and the authentication of commercial food products labelled as Agordino. The commercial varieties used as reference standards were chosen among those most grown by farmers in the Veneto region and were used to assess the intra- and inter-population genetic variation and to reconstruct the genetic structure of the population as whole.

For the purposes of this analysis, an initial screening was performed by using the 14 SSR markers selected from the work of Varshney *et al.* [8]. SSR markers were amplified and their PIC values assessed using a subset of samples belonging to both the local and the commercial varieties. Following this initial screening, the selection of the most suitable SSR markers was based on the following criteria: i) genetic mapping of the marker loci in different linkage groups, ii) lack of non-specific amplicons, iii) electropherograms showing unambiguous peaks, iv) ability to be amplified in multiplex reactions, v) absence of null alleles. As a result, seven markers were selected for the multilocus genotyping: Bmag0872, Bmag0125, EBmac0871, Bmag0808, EBmatc0003, Bmac0727 and Bmag0321.

The Agordino landrace scored a relatively high number of marker alleles among its individuals and the population on the whole proved to be composed of a mixture of closely related genotypes belonging to distinct pure lines, although homozygous for different marker alleles at the genomic loci investigated (*i.e.* the degree of homozygosity was 100 % for all individuals). In fact, most likely these pure lines, although cultivated one next to the other, remain reproductively independent and

therefore provide phenotypically homogenous progenies while maintaining substantial genetic variability among the pure lines.

The molecular marker-based genotyping is known to be accurate and effective in identifying the genetic signatures of crop varieties. In previous population genetic studies on barley, 10 to 20 SSR marker loci were analysed [20–23], whereas in this study the number of SSR marker loci was progressively reduced from 14 to 7 and then to four for genetic traceability purposes. Although it is difficult to define an ideal or standard number of microsatellite regions to test in a given species, the number of microsatellite markers depends on several factors, including the polymorphism information content of each target locus, the number of plant accessions and type of plant populations, and especially the goals of the genetic characterization. In this study, only one highly representative and informative SSR marker was selected for each of the linkage groups of barley and almost all individuals collected as Agordino could be unambiguously assigned to the landrace, on the basis of multilocus genotypes. Our findings demonstrate that the chosen SSR markers were highly efficient in revealing the genetic diversity and in assessing the genetic identity of the germplasm of barley accessions.

With the UPGMA grouping analyses, a reliable representation of the results was provided: the accessions were all clustered, with the strongest bootstrap support (100 %), into two genetically distinct subgroups, which were formed by the Agordino landrace individuals and by the commercial lines, respectively. The clear separation of the Agordino population into four or more different subgroups is consistent with a self-pollinating species likely composed of a mixture of genotypically similar but reproductively independent pure lines. Only a few samples of Agordino landrace (*i.e.* OA30, OA46, OA50 and OA52) scored low levels of genetic similarity compared with the rest of the individuals. Notably, these deviant genotypes of Agordino were tightly grouped along with some commercial varieties (see, for instance, Figures 1 and 2). This result suggests that seeds from commercial varieties might have been accidentally bulked with Agordino materials. Our

findings also support the possibility that some individual plants of the local variety might have been used to breed commercial varieties. For instance, the high mean genetic similarity value (>86 %) scored by the pure line Alba with Agordino individuals indicates that local materials might have contributed to the development of this commercial variety. Very scarce information is available for this variety. It is known that Alba is a variety bred for the ability of its malted grains for brewing and it is cultivated in Italy since the 1970s [24]. Actually this variety is not officially recorded in the national register of cultivated varieties and, to the best of our knowledge, there are no records about its ancestry.

Consistent with these findings, the detailed investigation using STRUCTURE v. 2.2 showed that the 95 samples were separated into two genetically distinct subgroups (K=2): most of the 60 analysed individuals of Agordino and the 21 commercial pure lines. Additionally, most of the samples (87 of 95) were assigned to either of the groups with accession scores for individual membership that were almost always higher than 95 % (with the exception of the four samples with values that ranged from 78 to 91 %). Based on this analysis, the Agordino population was homogeneous for high ancestry assignation. It is worth noting that the four Agordino individuals showing a membership higher than 95 % to the commercial group (*i.e.* OA30, OA46, OA50 and OA52) exhibited also the lowest mean genetic similarity values with the Agordino landrace, grouping closely to the commercial lines (see, for instance, Figures 1 and 2). These findings supported the hypothesis that seeds from commercial varieties might have been accidentally bulked with Agordino materials.

Four samples, *i.e.* OA20, OA38, OA44 and the pure line Alba, had admixed ancestry, with membership values that varied from 43 to 65 %. The admixed ancestry of the Alba variety (43 %) suggested that this commercial line was developed from an ancestor of the local Agordino variety. By contrast, for the admixed samples of Agordino landrace (*i.e.* OA20, OA38 and OA44), we

speculate that hybridization with commercial varieties could have occurred with important effects on the genetic structure of the landrace.

An attempt to increase the genetic distinctiveness and stability of this very old landrace still grown in Veneto has been done by taking advantage of the genetic structure data here presented. In particular, the samples characterized by a membership lower than 75 % (*i.e.* OA20, OA30, OA38, OA44, OA46, OA50, OA52) were removed and the population was multiplied using the seed set from the rest of individuals.

Concerning the need to implement a genetic authentication assay for the Agordino barley, the genotyping of all individuals allowed the identification of typical marker allele variants for this landrace. Based on these data, the most informative loci proved to be Bmag0872 (typical alleles D and F), Bmag0808 (typical alleles A, C and E), EBmatc0003 (typical allele A) and Bmag0125 (typical alleles A and D). These four SSR markers were successfully used not only to recognize and differentiate the local variety from other commercial varieties, but also to genetically trace its derived food products. Nevertheless, we do not exclude the possibility to investigate additional genomic loci to increase the number of Agordino-specific markers. The finding that typical marker alleles of Agordino could be unambiguously detected in experimental blends assembled by imposing up to tenfold dilutions of the target DNA (see Figure 5) indicates that the molecular identification assay is suitable to trace DNA from Agordino barley in a wide range of applications and products.

Since mislabeling is recognized as a significantly growing problem in food chains, our data indicate that a few chosen SSR markers are sufficient, when used in combination, to discriminate food products entirely prepared with traditional/local varieties from the ones blended with commercial ones. Furthermore, based on these data, we are planning to further investigate whether it is possible to correlate the intensity of amplified marker allele variants with the relative proportion of target varieties in a mixture.

Molecular data deriving from the amplification of genomic DNA extracted from two commercial Agordino-based food products revealed high levels of complexity as expected considering the high degree of genetic diversity found within the Agordino gene pool. Indeed, while several peaks were detected at each locus, only some of them (see Figure 6) could be unambiguously associated with the Agordino gene pool. Hence, crosschecks of data originating from the analysis of multiple SSR marker loci may be recommended, if not necessary, to efficiently capture the target marker alleles within the Agordino gene pool.

Based on our findings, the food products examined here seem to be entirely prepared with accessions of barley ascribable to the Agordino landrace, as stated in the food label. Despite the common origin of the two Agordino food products, significant differences in terms of peak intensities were found between them, suggesting that multiple genotypes and variable dosages of the same genotypes may have been adopted for the preparation of the two commercial products. Since qPCR is a more efficient approach to estimate the quantitative contribution of each genotype in the preparation of the commercial products, it could be employed as a future perspective.

Conclusions

Landraces that remain locally dominant in regional agriculture systems and, in particular, crop plant populations that have characterized a territory for a long time, must be safeguarded, and the cultural heritage and the landscape linked to the crop must be preserved. This study is the first detailed genetic characterization and description of the Agordino landrace of barley, an ancient and traditional variety widespread across the Belluno provincial area (Veneto, Italy) and widely used for the preparation of premium products sold locally.

An important result of this study is that the old Italian local variety of barley cultivated in the region of Agordo is represented by a small and well-defined group of genetically identifiable lines, well-separated and genetically differentiated from the other commercial varieties of barley cultivated in that region. This finding is extremely relevant for the development of a robust, fast and affordable

molecular identification assay based on a set of informative microsatellite loci showing marker alleles typical of the Agordino, hence useful for authenticating food derivatives and preventing food mislabelling.

Acknowledgements

This work was supported by the project ‘PROGRAMMA BIO.NET, rete regionale per la conservazione e caratterizzazione della biodiversità di interesse agrario - Gruppo di lavoro cerealicolo (WP5)’ funded by Programma di Sviluppo Rurale per il Veneto 2007-2013, Misura 214H, coordinated by Dr. Maurizio Arduin. The authors wish to thank Mr. Mirko Volpato and Mr. Stefano Cherubin for technical assistance with DNA extraction and PCR amplification experiments. Thanks are also due to Dr. Pino Silvio (Istituto di Genetica e Sperimentazione Agraria N. Strampelli, Lonigo, Italy) for providing seed stocks.

References

1. Barcaccia G, Molinari L, Porfiri O, Veronesi F. Molecular characterization of emmer (*Triticum dicoccon* Schrank) Italian landraces. *Genet Resour Crop Evol.* 2002;49(4):415–26.
2. FAO. Food and Agriculture Organization of the United Nations: Value of Agricultural Production [Internet]. 2016 [cited 2016 May 30]. Available from: <http://faostat3.fao.org/download/Q/QV/E>
3. Maresio Bazolle A. Il possidente bellunese. Maresio Bazolle A, Perco D, editors. Feltre (Italy): Comunità Montana Feltrina; 1986. 960 p.
4. Frank R, Jahn F, Barbiani G. Traditional plant from Carinzia, Friuli Venezia Giulia and Veneto. Regional agency for rural development - ERSA; 2012.
5. Scarano D, Rao R. DNA markers for food products authentication. *Diversity.* 2014;6(3):579–96.
6. Barcaccia G, Volpato M, Gentili R, Abeli T, Galla G, Orsenigo S, et al. Genetic identity of common buckwheat (*Fagopyrum esculentum* Moench) landraces locally cultivated in the Alps. *Genet Resour Crop Evol.* 2016;63(4):639–51.
7. Schuelke M. An economic method for the fluorescent labeling of PCR fragments A poor man's approach to genotyping for research and high-throughput diagnostics . *Nat Biotechnol.* 2000;18:233–4.
8. Varshney RK, Marcel TC, Ramsay L, Russell J, Röder MS, Stein N, et al. A high density barley microsatellite consensus map with 775 SSR loci. *Theor Appl Genet.* 2007 Apr;114(6):1091–103.
9. Applied Biosystem. Peak Scanner™ Software. Waltham, MA, USA: Thermo Fisher Scientific; 2006.
10. Nagy S, Poczai P, Cernák I, Gorji AM, Hegedus G, Taller J. PICcalc: An online program to calculate polymorphic information content for molecular genetic studies. *Biochem Genet.* 2012;50(9–10):670–2.
11. Yeh F, Yang R, Boyle T, Ye Z, Mao J. POPGENE: the user friendly shareware for population genetic analysis. Edmonton, Canada: University of Alberta; 1997.
12. Kimura M, Crow J. The number of alleles that can be maintained in a finite population. *Genetics.* 1964;49:725–38.
13. Nei M. Analysis of gene diversity in subdivided populations. *Proc Nat Acad Sci.* 1973;70(12):3321–3.
14. Lewontin R. The genetic basis of evolutionary change. New York, NY, USA: Columbia University Press; 1974.
15. Rohlf JF. NTSYS-pc: numerical taxonomy and multivariate analysis system, version 1.80. New York: Applied Biostatistics Inc; 1993.
16. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics.* 2003;164(4):1567–87.
17. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* 2000;155(2):945–59.
18. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol Ecol.* 2005;14(8):2611–20.
19. Doyle J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull.* 1987;19:11–5.
20. Jilal A, Grando S, Henry RJ, Lee LS, Rice N, Hill H, et al. Genetic diversity of ICARDA's worldwide barley landrace collection. *Genet Resour Crop Evol.* 2008;55(8):1221–30.
21. Chen ZW, Lu RJ, Zou L, Du ZZ, Gao RH, He T, et al. Genetic diversity analysis of barley landraces and cultivars in the Shanghai region of China. *Genet Mol Res.* 2012;11(1):644–50.

22. Khodayari H, Saeidi H, Roofigar AA, Rahiminejad MR, Pourkheirandish M, Komatsuda T. Genetic Diversity of Cultivated Barley Landraces in Iran Measured Using Microsatellites. *Int J Biosci Biochem Bioinforma.* 2012;2(4):287–90.
23. Bellucci E, Bitocchi E, Rau D, Nanni L, Ferradini N, Giardini A, et al. Population structure of barley landrace populations and gene-flow with modern varieties. *PLoS One.* 2013;8(12).
24. Villavecchia G, Eigenmann G. *Nuovo dizionario di merceologia e chimica applicata.* Milano, Italy: Hoepli; 1975.

Chapter III

Venetian local corn (*Zea mays* L.) germplasm: disclosing the genetic anatomy of old landraces suited for typical cornmeal mush production

Abstract

Due to growing concern for the genetic erosion of local varieties, the four main corn landraces grown in Veneto (Italy) -Sponcio, Marano, Biancoperla and Rosso Piave- were characterized in this work. A total of 197 phenotypically representative plants from these four landraces were genotyped at 10 SSR marker loci, which were regularly distributed across the 10 linkage groups and had an average polymorphism information content (PIC) of 0.5. In the population structure analysis based on this marker set, 144 individuals were assigned with strong ancestry association (90%) to four (K=4) distinct clusters, corresponding to the number of local varieties used in this study. The remaining 53 samples, mainly from Sponcio and Marano, showed admixed ancestry. Among all possible pairwise comparisons among landraces, these two varieties exhibited the highest mean genetic similarity (approximately 67%), as graphically confirmed through ordination based on PCoA and a UPGMA tree. These findings support the hypothesis of direct gene flow between Sponcio and Marano, highly promoted by their geographical proximity and overlapping cultivation areas. Conversely, consistent with a production mainly confined to the eastern area of the Veneto region, Rosso Piave showed the lowest similarity (<60%) to the other three landraces and firmly grouped (average membership=89%) in a separate cluster, forming a genetically distinguishable gene pool. Finally, although Biancoperla was represented at K=4 by a unique group with individual memberships higher than 80% in almost all cases (57 of 62), when analyzed with an additional level of population structure (K=6) it appeared to be entirely (100%) constituted by admixed individuals. This suggests that the current population could be the result of repeated hybridization events between the two accessions currently bred in Veneto, ITA0340323 and ITA0340324. The genetic characterization of these heritage landraces should prove very useful for monitoring and preventing further genetic erosion and genetic introgression, thus preserving their gene pools, phenotypic identities and qualitative traits for the future.

Keywords: microsatellite, genetic erosion, local varieties, maize, SSR; biodiversity; Veneto Region

Introduction

The concepts of genetic erosion and conservation of plant genetic resources are rooted in the first decade of the twentieth century. Since then, several authors have warned of the consequences of the reduction of genetic variability in crop species mainly due to the dramatic loss of traditional landraces [1–3]. A landrace is an ancient population of a cultivated crop plant that has become adapted to the local conditions and to the agronomic practices of farmers. Most frequently, landraces are characterized by high diversity and thus provide a valuable source for potentially useful traits and an irreplaceable bank of co-adapted genotypes [4]. In practical terms, genetic diversity allows farmers and plant breeders to adapt a crop to heterogeneous and changing environments by, for example, providing it with resistance to pests and diseases [5].

Since modern, highly productive cultivars are irreversibly replacing many traditional varieties, the first priority is to arrest this loss of genetic diversity.

Over the last decade, the rediscovery of local and traditional food products in the market has strengthened interest in local varieties. A fascinating case study is represented by ‘polenta’, a traditional dish of the Italian cuisine, and by the four main corn landraces grown in Veneto (Italy) used for its production: Sponcio, Marano, Biancoperla and Rosso Piave. In the last few years, the demand of ‘polenta’ from local varieties, has shown a steady increase due to the deeper attention that consumers pay to the autochthonous, locally cultivated crops, usually grown according to low-input agronomic practices, and to their consciousness towards the current dualism existing between conventional and novel foods [6]. In 2016 the total production of maize in Italy was approximately 6.84 million tonnes [7], and even if the amount destined to the human consumption is very small (few percentage points), the total market value of this product is estimated in millions of euro.

Assessing the genetic diversity and genetic structure of landraces could help to limit genetic erosion as well as to conserve landraces [8]. Several studies have been performed worldwide to assess the genetic diversity of local landraces of corn using molecular markers [9–12], but as far as we know,

only one has been devoted to local varieties of Italian corn [13]. The four main corn landraces grown in Veneto (Italy) and examined in this work, namely, Sponcio, Marano, Biancoperla and Rosso Piave, represent a case study.

Sponcio is an ancient corn variety grown by a consortium of 20 farmers in a small plot that covers approximately 13 hectares in the area of the Val Belluna, specifically in the towns of Feltre, Cesiomaggiore and Santa Giustina [14]. This landrace, distinguishable by its sharp kernels, seems to have been known since the sixteenth century under variants of the name, but the first concrete documentation of its existence is a nineteenth century manuscript [15]. By the 1950s, the production of Sponcio had been reduced, and it was confined to marginal areas. Thanks to a few farmers and millers, the original germplasm was carefully preserved and later used to restart the current production, according to strict sustainable and environmentally friendly regulations determining the stages of cultivation, drying, grinding and packing. The yellow flour is the main ingredient of '*polenta*', one of the most typical products of the Belluno cuisine.

An article dated 1939 reports that in 1890 Antonio Fioretti, a farmer from Marano Vicentino (Veneto, Italy), crossed two local varieties, Nostranino and Pignoletto d'Oro, and called this new hybrid Marano [16]. Although Marano was particularly esteemed during the '70s in Veneto and Friuli Venezia Giulia and widely employed to produce new hybrids (e.g., ITALO 225, ITALO 260 and ITALO 270) and pure lines (Cinquantino San Fermo and Cinquantino Bianchi), the cultivation of this local variety was progressively abandoned and replaced by more productive lines. Currently, it survives only in the area of Val Leogra, specifically in the towns of Marano Vicentino, Malo, Schio, San Vito di Leguzzano, Torrebelvicino, Valli del Pasubio, Santorso and Piovene Rocchette [17]. The '*polenta maranelo*', a typical dish of this area, is produced starting from the orange flour of this landrace.

A book published at the end of seventeenth century reports that a white '*sorgoturco*' [dialectal word referring to corn] was widespread in Veneto at that time, and that white variety probably represents

the ancestor of the current Biancoperla [18]. Several documented sources state that this landrace, which owes its name to the vitreous and pearly white color of its kernels, was widely grown (> 50,000 hectares) in the eastern part of Veneto and in Friuli Venezia Giulia in the first half of the twentieth century [19]. As with the aforementioned landraces, this local variety was progressively replaced by more profitable corn varieties from the USA immediately after the Second World War. Currently, thanks to a consortium of approximately 13 member producers promoting its conservation, this variety survives on less than 50 hectares in some rural areas of Vicenza, Treviso and the northern part of Padua district. It is strongly appreciated for the production of white '*polenta*' [20].

Little is known about the fourth landrace, Rosso Piave. Miniscalco [19] reports that, unlike the other three local varieties, this landrace was rarely grown even in the past, since its color permanently soiled mills. Today, it is grown mainly in the Venice area in the towns of Musile di Piave, Fossalta di Piave, Noventa di Piave and San Donà di Piave. Its peculiar burgundy color, which also characterizes its '*polenta*' - the main derivative product of this landrace - comes from the presence of anthocyanins that have been recently recognized as compounds able to reduce the risk of myocardial infarction [21].

In this study, the genetic diversity of the four main corn landraces in Veneto was assessed by means of simple sequence repeat (SSR) markers. The assembled molecular data were used to evaluate their population genetic structure and their genetic relationships. The characterization of these old local varieties supports a more general discussion of possibilities for avoiding genetic erosion, promoting and safeguarding local populations, thereby maintaining stable seed yields, and preserving phenotype and qualitative identity.

Materials and Methods

Plant material and genomic DNA isolation

Four different Venetian Institutes for Agricultural Research kindly donated the corn samples used in the present study. The germplasm collection conserved in each institute was originally constituted combining hundreds of kernels from as many ears selected on the basis of their morphology. Marano seeds were provided by the “N. Strampelli Institute” (Lonigo, VI, Italy). Sponcio seeds were obtained from the “Antonio della Lucia Institute” (Feltre, BL, Italy). The “D. Sartor Institute” (Castelfranco Veneto, TV, Italy) and Veneto Agricoltura (Legnaro, PD, Italy) supplied Biancoperla and Rosso Piave seeds, respectively.

For germination, 40 to 70 seeds of each variety were placed in Petri dishes on two layers of filter paper moistened with water. After fifteen days of incubation, a total of 197 seedlings (64 Marano, 32 Sponcio, 62 Biancoperla and 39 Rosso Piave) were collected and used for the analyses described below. The inbred line B73 was used as tester.

Then, 100 mg of fresh leaf tissue was used to isolate genomic DNA using a DNeasy plant kit (Qiagen, Valencia, CA), following the procedure provided by the suppliers. Electrophoresis on an 0.8% agarose/1× TAE gel containing 1× Sybr Safe DNA stain (Life Technologies, Carlsbad, CA) allowed estimation of the integrity of extracted DNA samples. The purity and quantity of DNA extracts were evaluated with a NanoDrop 2000c UV-Vis spectrophotometer (Thermo Scientific, Pittsburgh, PA).

Analysis of SSR markers

PCR amplifications were performed using the M13-tailed SSR method described by Schuelke [22], with some minor modification. Briefly, the amplification procedure is based on a three-primer system consisting of a specific SSR-targeting forward primer with a 5'-M13 tail, a specific SSR-targeting reverse primer and an M13-labelled primer (5'-TTGTAAAACGACGGCCAGT-3'). The set of 10 SSR marker loci investigated in this study was obtained from Register *et al.* [23] and,

based on the highest Polymorphic Information Content (PIC) values, one SSR marker per linkage group was selected (Table 1).

Table 1. List of SSR loci selected from Register *et al.* [23] for use in this study. For each microsatellite locus, linkage group, locus ID, motif, amplicon size in bp, primer sequences used to amplify the region, melting temperature and polymorphism information content (PIC) coefficient related to the previously mentioned study are shown

Linkage group	Locus ID	Motif	Size (bp)	Primer	T _m (°C)	PIC
1	phi056	GCC	103-112	M13-ACGCCCAGATCTGTTTCCTTCTC ATGGCGGCAGGCCGATTGTT	63	0.67
2	phi127	AGAC	129-145	M13-ATATGCATTGCCTGGAAGTGGAAAGGA AATCAAACACGCCTCCCGAGTGT	62	0.70
3	phi073	CAG	107-116	M13-TTACTCCTATCCACTGCGGCCTGGAC GCGGCATCCCGTACAGCTTCAGA	69	0.65
4	phi076	GAGCGG	182-192	M13-TTCTTCCGCGGCTTCAATTTGACC GCATCAGGACCCGCAGAGTC	61	0.65
5	phi024	CCT	183-195	ACTGTTCCACCAAACCAAGCCGAGA M13-AGTAGGGGTTGGGGATCTCCTCC	69	0.69
6	phi031	GTAC	174-177	M13-GCAACAGGTTACATGAGCTGACGA CCAGCGTGCTGTTCCAGTAGTT	66	0.57
7	phi057	GCC	211-215	M13-CTCATCAGTGCCGTCGTCCAT CAGTCGCAAGAAACCGTTGCC	66	0.61
8	umc1075	ATTGC	156-166	M13-GAGAGATGACAGACACATCCTTGG ACATTTATGATACCGGGAGTTGGA	57	0.69
9	phi016	GGT	173-176	M13-TTCCATCATTGATCCGGGTGTCG AAGGAGCAACATCCCATCCAGGAA	60	0.52
10	phi084	GAA	174-178	M13-AGAAGGAATCCGATCCATCCAAGC CACCCGTA CTTGAGGAAAACCC	59	0.49

The PCR reaction consisted of a 20 µl final volume containing 1× NH₄ Reaction Buffer, 3 mM MgCl₂, 1 IU of BioTAQ™ DNA polymerase (Bioline, London, UK), 0.25 mM each dNTPs, 0.25 µM tailed forward primer, 0.75 µM reverse primer, 0.5 µM M13-labelled primer (Invitrogen, Carlsbad CA), 20 ng of DNA, and dH₂O up to the final volume. Amplifications were performed in a 96-well plate using a 9600 thermal cycler (Applied Biosystems, Carlsbad CA). The following thermal conditions were used: 5 min at 94°C for the initial denaturing; 5 cycles at 94°C for 30 s, at 62°C for 30 s decreasing by 0.8°C with each cycle, and at 72°C for 45 s; and 35 cycles at 94°C for 30 s, 58°C for 30 s and 72°C for 45 s. A final extension at 72°C for 10 min terminated the reaction to fill in any protruding ends of the newly synthesized strands. Capillary electrophoresis with an ABI PRISM 3130xl Genetic Analyzer, adopting LIZ500 as molecular weight standard, was used to

assess the PCR products. The size of each peak was determined using Peak Scanner software 1.0 (Applied Biosystems).

Marker data analysis

PIC values were calculated with PICcalc software [24] to estimate the marker allele variation in microsatellite loci in the 197 corn individuals. GenA1EX v.6.5 [25] and POPGENE v.1.32 [26] software were used to estimate the number of observed alleles (N_a), number of effective alleles (N_e), Shannon's information index of genetic diversity (I), observed (H_o) and expected (H_e) heterozygosity according to Nei [27]. The presence of private alleles in each population and the occurrence of locally common alleles, defining as 'locally common' those alleles with a frequency higher than 5% in a local population and occurring in less than 25% of all populations examined [28], were also considered.

F-statistics were calculated according to Wright [29] to investigate the variance of heterozygosity in our population at different levels of population structure (i.e., individual, subpopulation and population levels). Inbreeding coefficients were computed to measure the deficiency (positive values) or excess (negative values) of heterozygotes for each assessed microsatellite marker and to assess hierarchical organization of sample individuals. Similarly, inbreeding coefficients were calculated at multilocus level in order to estimate the genetic effect of total population subdivision as proportional reduction in overall heterozygosity due to variation in SSR allele frequencies among different subpopulations [30]. Finally, gene flow (N_m) was calculated from F_{ST} .

Genetic similarity (GS) was calculated in all possible pairwise comparisons of individuals by applying the simple matching coefficient [31]. A principal coordinates analysis (PCoA) was applied to compute the first two principal components of the similarity data matrix. All analyses and calculations were conducted using NTSYS-pc v. 2.21q [31].

Pairwise GS values were also used to compute a dendrogram, which was constructed using the unweighted pair-group arithmetic average (UPGMA) method and PAST software v. 3.14 [32] with 1000 bootstrap repetitions.

The genetic structure of the four landraces was modeled using a Bayesian clustering algorithm implemented in STRUCTURE v. 2.2 [33]. Since no prior knowledge about the origin of the populations under study was available, the *admixture model* was used; a *correlated allele frequencies* model was selected since it guarantees that a previously undetected correlation will be identified, without affecting the results if no such correlation exists [34]. Ten replicate simulations were conducted for each value of K, with the number of founding groups ranging from 2 to 22, using a burn-in of 2×10^5 and a final run of 10^6 MCMC steps. The method described by Evanno *et al.* [35] was used to evaluate the most likely estimation of K. Estimates of membership were plotted as a histogram using an Excel spreadsheet.

Results

Descriptive statistics of SSR marker loci

All SSR loci were determined to be polymorphic (Table 2). PIC values were considered to estimate the ability of each locus to discriminate among different genotypes, and the selected SSR loci had a mean PIC of 0.50 with a minimum of 0.32 (phi084) and a maximum of 0.71 (umc1075).

Thirty-six alleles were detected across four populations with an average number of observed alleles (N_a) of 3.60, ranging from two (phi084, phi031 and phi016) to six (phi127). N_e ranged from 1.68 (phi084) to 4.02 (umc1075, Table 2).

SSR loci were highly polymorphic within each landrace, except for phi084, which was monomorphic in the Marano landrace for a 177 bp marker allele. The same allele was also the most common one overall, being detected in 141 out 197 samples (71.68%, Supplementary Figure S1). Six alleles of the 36 were private to single populations. Specifically, Sponcio showed two private alleles, one at the phi057 locus and one at the phi056 locus, while Rosso Piave showed four

different allelic variants never detected in the other landraces, three at the phi127 locus and one at the phi024 locus. The number of “locally common alleles” was always equal to 0 in every population.

Over all loci, the observed (H_o) and expected (H_e) heterozygosity estimates were, on average, equal to 0.43 (± 0.12) and 0.58 (± 0.12 , Table 2). The same indexes calculated within each landrace were, on average, equal to 0.43 (± 0.04) and 0.48 (± 0.04), respectively (Table 2).

Table 2. Genetic parameters with respect to the SSR markers and to the four landraces object of this study. Average number of observed alleles (N_a), effective number of alleles (N_e) per locus, polymorphism information content (PIC), estimates of Shannon’s information index of genetic diversity (I), observed heterozygosity (H_o) and unbiased Nei’s genetic diversity equivalent to the expected heterozygosity (H_e) are shown. Wright’s inbreeding coefficients F_{IS} , F_{IT} and F_{ST} and gene flow (N_m) estimates are also indicated

Locus	N_a	N_e	PIC	I	H_o	H_e	F_{IS}	F_{IT}	F_{ST}	N_m
phi024	5.00	3.54	0.66	1.32	0.59	0.72	0.07	0.19	0.13	1.69
phi127	6.00	2.24	0.47	0.95	0.43	0.55	0.12	0.22	0.11	1.97
phi084	2.00	1.68	0.32	0.60	0.29	0.41	-0.03	0.26	0.28	0.65
phi076	3.00	2.95	0.59	1.09	0.48	0.66	0.14	0.24	0.12	1.84
phi031	2.00	1.75	0.34	0.62	0.39	0.43	0.04	0.09	0.05	4.87
phi057	4.00	2.72	0.55	1.07	0.48	0.63	0.01	0.27	0.26	0.70
phi056	4.00	1.86	0.38	0.77	0.29	0.46	-0.08	0.37	0.42	0.34
phi073	3.00	2.94	0.59	1.09	0.55	0.66	0.01	0.17	0.17	1.24
phi016	2.00	1.00	0.37	0.69	0.250	0.50	0.34	0.47	0.19	1.04
umc1075	5.00	4.02	0.71	1.47	0.53	0.75	0.16	0.26	0.12	1.84
All loci	3.60	2.57	0.50	0.97	0.43	0.58	0.08	0.25	0.18	1.62
St.dev	1.43	0.80	0.14	0.30	0.12	0.12	0.04	0.03	0.03	0.40
Sponcio	3.00	2.09	na	0.80	0.46	0.49	0.06	0.21	0.16	1.34
Marano	2.50	2.08	na	0.74	0.46	0.47	0.02	0.21	0.19	1.04
Biancoperla	2.50	1.90	na	0.69	0.37	0.44	0.15	0.35	0.24	0.78
Rosso Piave	3.30	2.30	na	0.89	0.45	0.53	0.16	0.23	0.08	3.01
All landraces	2.82	2.09	na	0.80	0.43	0.48	0.10	0.25	0.17	1.54
St.dev	0.39	0.16	na	0.09	0.04	0.04	0.07	0.07	0.07	1.00

Shannon index (I) was used to characterize population diversity and was found to be, on average, equal to 0.97 (± 0.30) over all loci and 0.80 (± 0.09) within landraces. The inbreeding coefficient (F_{IS}) had an average value of 0.08 (± 0.04) for SSR loci. Finally, F_{IT} and F_{ST} were, on average, both positive and equal to 0.25 (± 0.03) and 0.18 (± 0.03), respectively, while the gene flow (N_m), calculated from F_{ST} , was equal to 1.62 (± 0.40).

The same F-statistics were then applied to each landrace to assess the genetic effects of total population subdivision as proportional reduction in overall heterozygosity due to variation in SSR

allele frequencies among landraces (Table 2). Overall, Wright's inbreeding coefficients F_{IS} and F_{IT} scored positive values, revealing a general deficiency of heterozygotes across individual accessions and landraces. As displayed in Table 2, reduction of heterozygosity was marked for the two landraces Rosso Piave ($F_{IS}=0.16$) and Biancoperla ($F_{IS}=0.15$), whereas it was minimal for the two landraces Sponcio ($F_{IS}=0.06$) and Marano ($F_{IS}=0.02$). Interestingly, the variation observed in our estimates of F_{IT} was much lower, as this parameter was on average equal to 0.25, ranging from 0.21 (Sponcio and Marano) to 0.35 (Biancoperla). As displayed in Table 2, F_{ST} was, on average, equal to 0.17, ranging from 0.08 (Rosso Piave) to 0.24 (Biancoperla, Table 2). Altogether, these data suggest that the proportion of genetic variation found among landraces was relatively low (17% on average) and to some extent variable across landraces (8% to 24%).

Genetic diversity and cluster analysis

Within the genetic similarity (GS) matrix calculated for all possible pairwise comparisons among the 197 DNA genotypes, Rohlf's index ranged from 29.01% (between MAR_128 and RSM_25) to 97.24% (between MAR_129 and SPO_32). When calculated within each landrace, this index varied, on average, from 69.98% ($\pm 10.55\%$) within the Rosso Piave population to 78.12% ($\pm 7.60\%$) within the Biancoperla population. In pairwise comparisons between varieties, Marano and Biancoperla showed the lowest average value ($59.10 \pm 7.14\%$) while Marano and Sponcio exhibited the highest one ($67.23 \pm 7.93\%$, Supplementary Figure S2).

Principal coordinate analysis (PCoA) showed that most of the samples were clustered into four major groups (Figure 1).

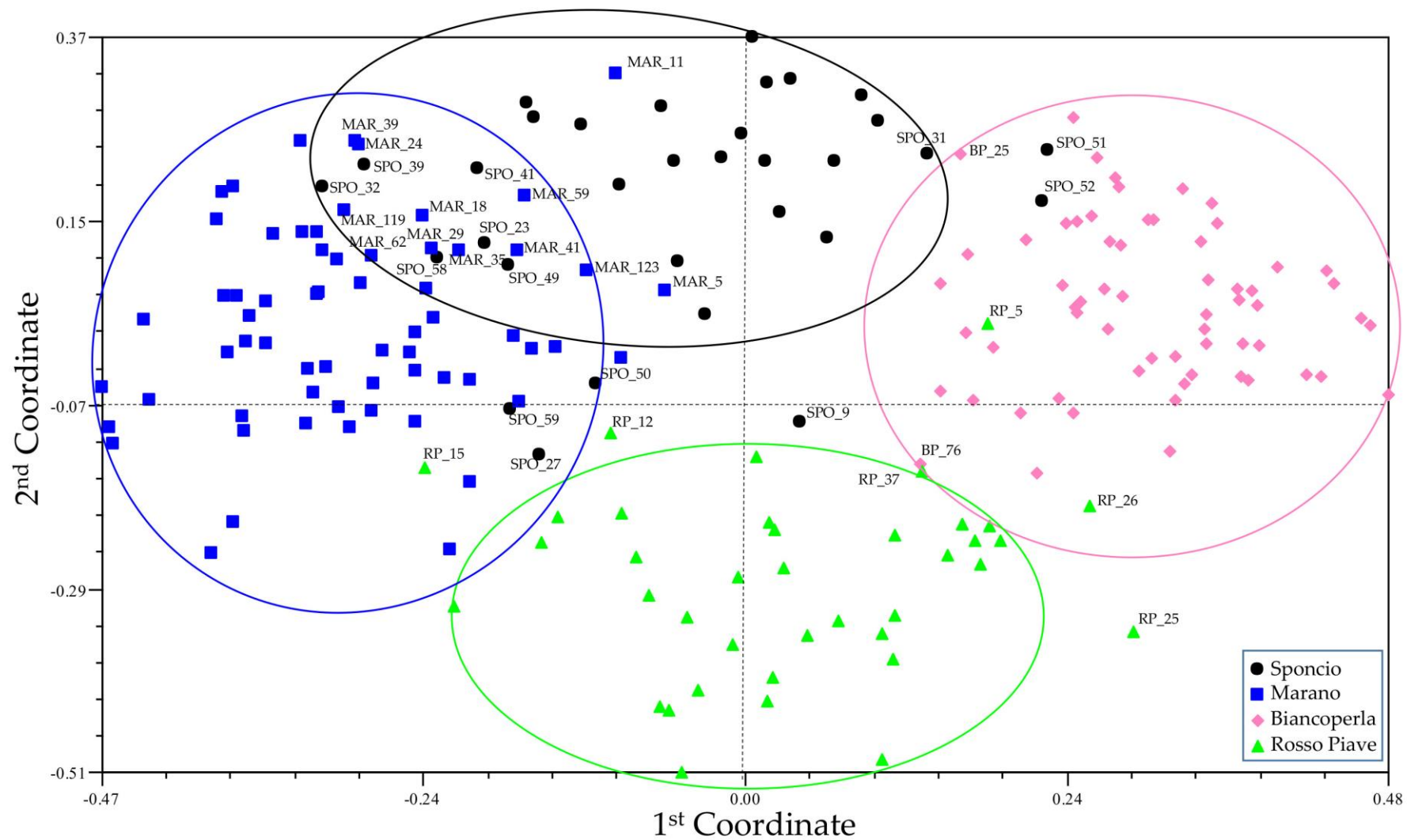


Figure 1. Two-dimensional centroids derived from genetic similarity estimates computed among the 197 accessions in all possible pairwise comparisons using the SSR marker data set. Only the names of those genotypes with unclear membership to one of the four subgroups are reported

The first principal coordinate accounted for 14.66% of the total variation and clearly separated Marano from Biancoperla, whereas the second principal coordinate accounted for 7.38% of the total variation and separated Sponcio from Rosso Piave. Biancoperla and Rosso Piave were firmly grouped in two distinct clusters and only few individuals, namely, RSM_5, RSM_12, RSM_15, RSM_25 and RSM_26, were partially separated from the rest of their landrace. This finding is also supported by low mean genetic similarity values (always lower than 69.00%) calculated through pairwise comparisons between these five samples and the Rosso Piave collection as a whole. The PCoA analysis further underlines the existence of some overlaps between the Marano and Sponcio clusters, a scenario that is mirrored by the high mean similarity value (67.23%) calculated at all loci between these two populations.

UPGMA cluster analyses revealed marked differentiation of the four local varieties. Using this approach, it was possible to distinguish three main clusters of individuals that were firmly supported by a bootstrap value of 100% (Figure 2).

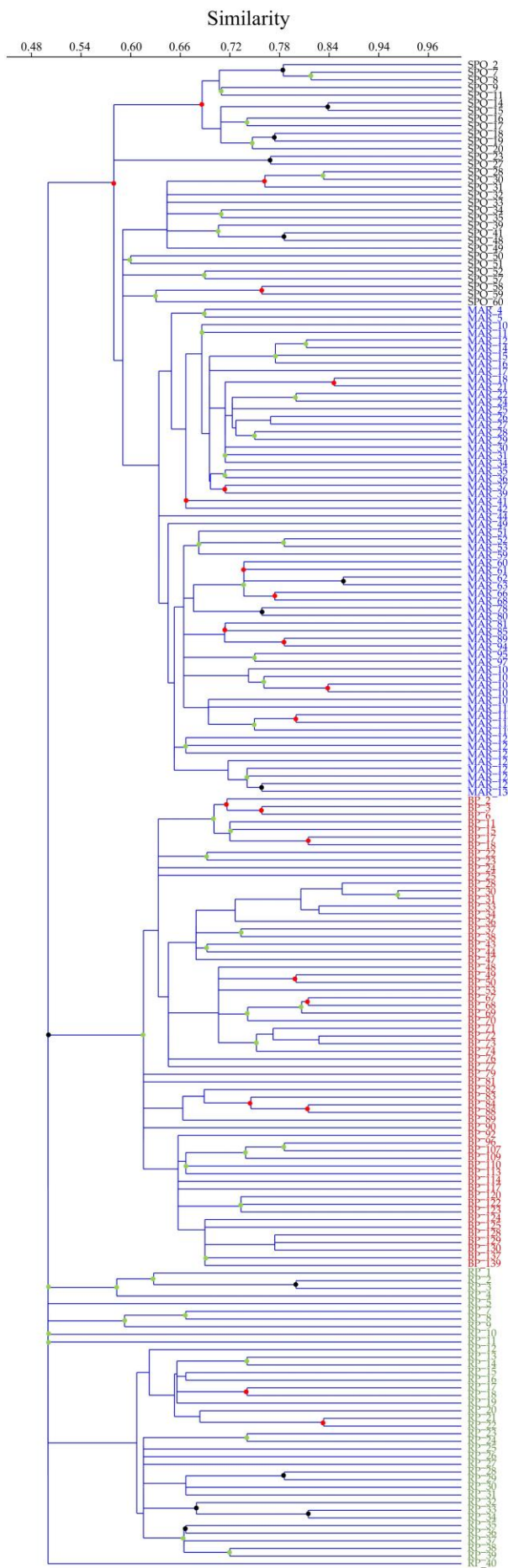


Figure 2. Constrained UPGMA tree of genetic similarity estimates computed among pairwise comparisons of corn accessions using the SSR marker data set, with nodes supported by bootstrap values. Black circle: bootstrap values $\geq 90\%$; red circle: $70\% \leq$ bootstrap values $< 90\%$; green circle: $50\% \leq$ bootstrap values $< 70\%$. The color scheme for the text is the same as for the symbols described in Figure 1. Black = Sponcio, blue = Marano, magenta = Biancoperla and green = Rosso Piave

The largest group consisted of two subgroups with bootstrap support of 76%, one including approximately 50% of the Sponcio population and the other one including the Marano population and the remaining Sponcio individuals. The second cluster included the entire Biancopera population, which further split into two subgroups (bootstrap support 52%). Finally, the third group represented most of the Rosso Piave individuals. The inbred line B73 was clustered separately from all landraces populations (data not shown).

Genetic structure analysis

STRUCTURE v2.2 [33] was used to investigate the genetic structure of the corn core collection. Following the procedure of Evanno *et al.* [35], a clear maximum for ΔK value at $K=4$ was found ($\Delta K=548.79$, Figure 3).

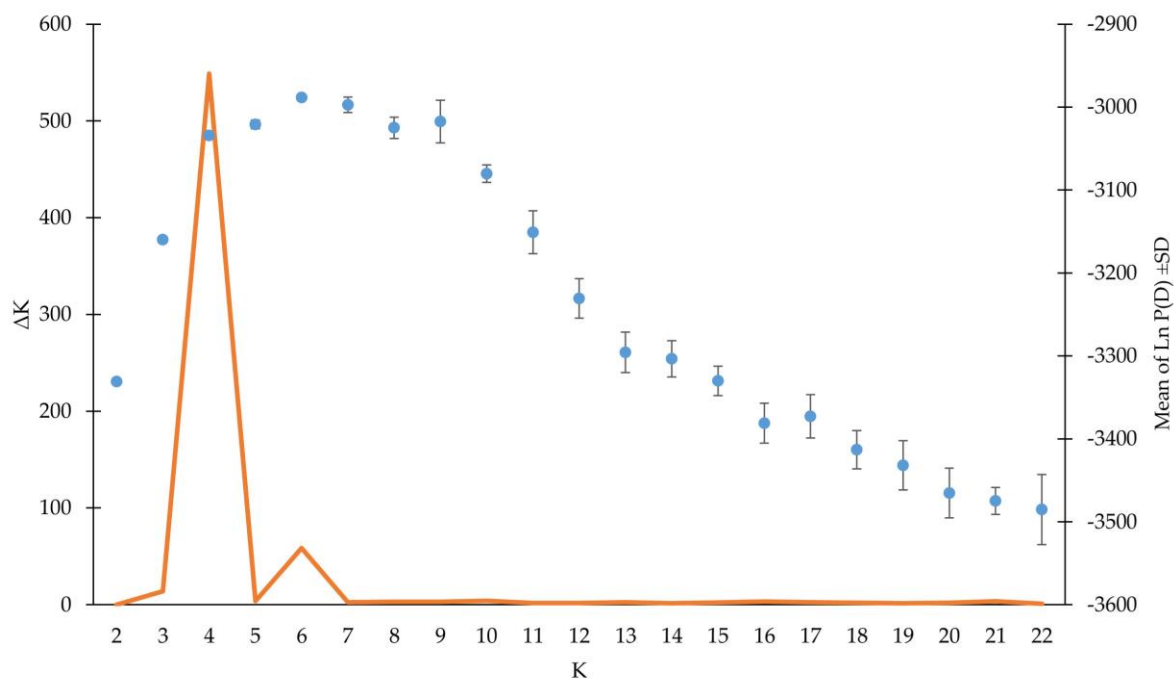


Figure 3. Definition of the number of ancestral corn populations based on the SSR marker dataset. Mean $\ln P(D) \pm SD$ over 10 runs is a function of K , as $L'(K) = \Delta \ln P(D)$. Mean ΔK is calculated as $|L''(K)| / (SD(L(K)))$ following Evanno *et al.* [35]. ΔK values are represented by the orange line, while the blue points indicate the mean $\ln P(D) \pm SD$ values

Since the ancestral population size $K=4$ also corresponds to the number of local varieties used in this study, it was considered the best estimate of the current population structure (Figure 4a). The 197 corn samples were grouped into four genetically distinct clusters. In this graphical

representation, each genotype is plotted as a vertical histogram divided into $K=4$ colored segments representing the estimated membership in each hypothesized ancestral genotype. The clustering of genotypes revealed that 144 of 197 samples showed strong ancestry association ($>90\%$). Almost all individuals from Biancoperla (94%) and Rosso Piave (90%) showed an individual membership to their respective founding groups higher than 80% while most of the admixed genotypes ($<80\%$ membership to a single ancestral genotype) were from Sponcio and Marano. Specifically, these two landraces included a substantial number of genotypes with admixed ancestry, namely, 11 genotypes for the variety from Val Belluna (32%) and 9 for the one from Marano Vicentino (15%). Most of the admixed genotypes (Figure 4a) originated in the overlapping region between these two clusters in the PCoA analysis (Figure 1, MAR_5, MAR_11, MAR_29, MAR_59, SPO_27, SPO_32, SPO_39, SPO_49 and SPO_58, SPO_59).

The second largest ΔK , at $K=6$ ($\Delta K=58.58$, Figure 3), revealed an additional level of population structure and allowed the clustering of all investigated genotypes into six additional subgroups. In this interpretative framework, all the Marano samples and most of the Sponcio population (91%) were organized into three main clusters. The first one included 21 individuals from Sponcio (membership $>50\%$), a second one grouped 39 Marano genotypes (membership $>50\%$) and a third cluster comprised most of the Sponcio and Marano samples that showed admixed ancestry at $K=4$ (Figure 4b). As already found for $K=4$, Rosso Piave continued to cluster apart, but all the individuals belonging to the Biancoperla landrace showed admixed ancestry from two different clusters (Figure 4b).

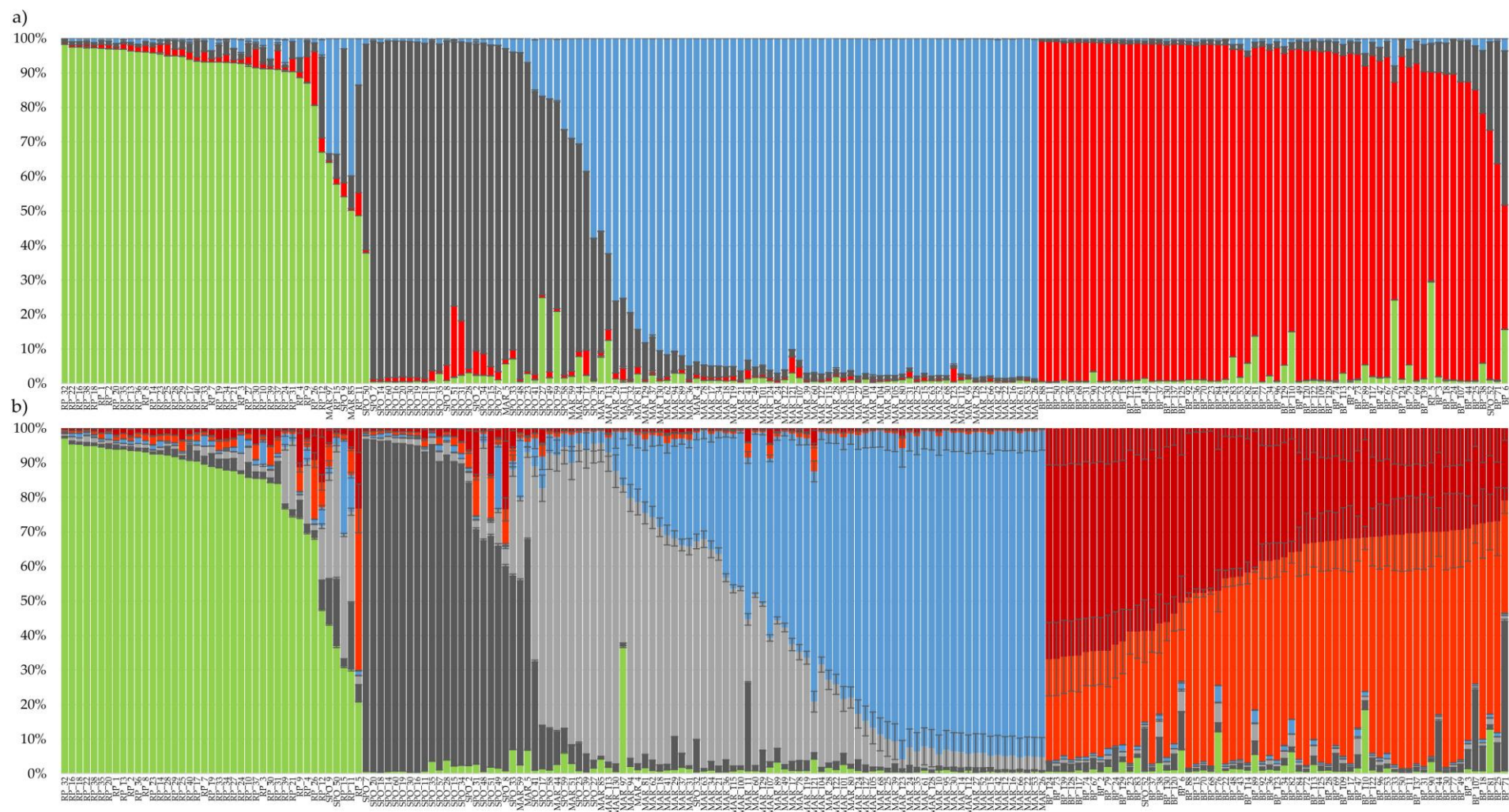


Figure 4. Population genetic structure of the four main corn landraces in Veneto (n=197) as estimated by STRUCTURE using the SSR marker data set. Each sample is represented by a vertical histogram partitioned into K=4 (panel a) or K=6 (panel b) colored segments that represent the estimated membership. The proportion of ancestry (%) is reported on the ordinate axis and the identification number of each accession is reported below each histogram. The color scheme for the figure is the same as for the symbols and the text described in Figures 1 and 2, respectively. Green = Rosso Piave, black = Sponcio, blue = Marano and magenta = Biancoperla. For K=6 two shades of red are used for the two clusters of Biancoperla and the third new cluster between Sponcio and Marano is marked in grey

Discussion

The Sponcio, Marano, Biancoperla and, to a lesser extent, Rosso Piave landraces of corn were abundantly grown in the past and characterized the Veneto region (Italy) for centuries, as reported by several documents [15,16,19]. Nevertheless, since the twentieth century, they have been progressively abandoned and replaced by more productive lines. Currently, they survive only in a few hectares, and extinction is becoming a real threat. To the best of our knowledge, this work represents the first attempt to describe the genetic diversity and structure of these four varieties.

For the purposes of this analysis, 10 microsatellite markers equally distributed into 10 linkage groups (Table 1) were chosen from Register *et al* [23] on the basis of their high PIC. The PIC values calculated for this dataset (Table 2) were slightly lower than those reported by Register *et al*. [23], probably because in the original work PIC values were obtained from an analysis of over 500 genotypes largely representative of the whole North American germplasm. Moreover, we cannot exclude that the different methods to run and screen PCR products could influence the detection of allelic variants. According to Botstein *et al*. [36], five of the selected SSR markers (phi024, phi076, phi057, phi073 and umc1075) would be considered highly informative ($PIC > 0.5$), while the other loci would be considered informative ($0.25 < PIC < 0.5$, Table 2). Interestingly, there was no direct correlation between population size and number of observed alleles (N_a) or number of effective alleles (N_e). In fact, the two populations numerically less represented, Sponcio ($N=32$) and Rosso Piave ($N=39$), showed the highest N_a ($N_a=3.00$ and $N_a=3.30$, respectively) and N_e values ($N_e=2.09$ and $N_e=2.30$, respectively). The number of effective alleles in a population is estimated from the gene diversity (i.e., $N_e=1/(1-H_e)$), and denotes the number of equally frequent alleles necessary to achieve a given level of gene diversity. The finding that the number of effective alleles indirectly correlates with the size of the assessed populations is consistent with a reduction of gene diversity within these populations (Figure 1). Furthermore, the observation that the difference between the number of observed alleles (N_a) and number of effective alleles (N_e) is higher in Sponcio and Rosso

Piave could indicate the presence of several low-frequency alleles in these landraces. Roughly speaking, without considering allele frequencies at this level of interpretation, a high number of observed alleles theoretically produces many genetically possible genotypes and, thus, high genetic diversity within the population. Accordingly, we observed that mean genetic similarity values scored within Rosso Piave (69.98%) and Sponcio (74.71%) were lower than those calculated within Marano and Biancoperla (75.40% and 78.12%, respectively). Of the 36 alleles, six appeared to be private to specific populations (Sponcio and Rosso Piave). SSR private alleles are recognized as an efficient food traceability tool since they can be assigned unambiguously to a specific variety. Recently, [37] a molecular system entirely based on private alleles to verify the genetic authenticity of food products deriving from an Italian barley landrace was developed. Unfortunately, all six private alleles observed in this study were present at very low frequencies ($<0.05\%$), so they could not be used, even in combination, for the same purpose. The large number of polymorphisms and the presence of both rare allele and alleles unshared with B73, potentially confirm that these four landraces could represent a valuable source of genetic variation and unique germplasm traits [13].

The fact that the overall mean observed heterozygosity ($H_o=0.43\pm 0.12$) for all loci was lower than expected ($H_e=0.58\pm 0.12$) suggests an excess of homozygosity in the core collection. This is further supported by the positive values of the individual inbreeding coefficients ($F_{is}=0.08\pm 0.04$ and $F_{it}=0.25\pm 0.03$, Table 2). The observed heterozygosity calculated within the four landraces was, on average, equal to 0.43 (± 0.04), consistent with the allogamous reproductive system of corn and with that reported in other works focused on corn landraces [11,12]. As found in the only other Italian work available on corn landraces [13], a deficiency of heterozygotes was observed for each local variety.

We are confident that the deficiency of heterozygosity is not correlated with the size of the assessed populations: the lowest value ($H_o=0.37\pm 0.20$) was recorded for Biancoperla, one of the two most numerically represented landraces ($N=62$) and *vice versa* the smallest group (Sponcio, $N=32$)

showed the highest value ($H_o=0.47\pm 0.15$). Moreover, we rule out the possibility that the cause of low levels of heterozygosity is ascribable to the limited number of plants from which the seeds sampled and analyzed in this study were originally selected by the institutes. In fact, germplasm collections were constituted combining hundreds of kernel corns from as many ears, carefully avoiding seeds from the same plant and ear. More likely, the repeated crosses of genetically similar individuals played a crucial role in the homozygosity excess showed by the loci investigated [38]. This could be the case when farmers select, every year, very small seed stocks based on an ‘ear ideotype,’ applying a strong selective pressure [13]. More in details, the traditional selection carried out annually by farmers is oriented to maintain i) the distinctive morphological traits of the landrace, ii) the peculiar qualitative characteristics of kernels used for ‘polenta’, and iii) the level of distinctiveness even when the pollen source is not controlled [6]. Biancoperla also scored the highest mean value of similarity (78.12%), providing a reasonable connection between the low level of heterozygosity and high genetic similarity calculated within each local variety.

Overall, Wright’s inbreeding coefficients F_{IS} and F_{IT} scored positive values, confirming a general deficiency of heterozygotes across individual accessions and landraces. Based on our marker set, the reduction of heterozygosity was higher for the two landraces Rosso Piave and Biancoperla, while it was relatively low for the two landraces Sponcio and Marano. Interestingly, F_{IT} estimates did not mirror the variation observed for F_{IS} , as the four cultivars displayed very similar values for this parameter. Estimates of F_{ST} varied considerably among the four landraces, indicating unbalanced contributions of the investigated populations to the total assayed genetic variation. Accordingly, our estimates of inbreeding coefficients suggest that these landraces are characterized by a relatively low degree of genetic differentiation, with approximately 17% of the genetic variation found among landraces (average $F_{ST}=0.17$) and approximately 83% of the total genetic variation expressed within landraces.

Based on a pairwise comparison among varieties, Sponcio and Marano exhibited the highest mean genetic similarity estimates (on average, 67.23%), as graphically confirmed through ordination analyses based on the definition of PCoA centroids (Figure 1) and construction of the UPGMA tree (Figure 2). Our findings support the hypothesis of marked gene flow between these two landraces, which could have been promoted by their geographical proximity and recently overlapping cultivation areas as a clear-cut distribution in the Veneto region has been progressively lost. To further corroborate this hypothesis, Marano and Sponcio revealed genetically differentiated populations for $K=4$ (Figure 4a) and subpopulations grouping individuals with admixed ancestry for $K=6$ (Figure 4b). Further analysis will be needed, and combining genetic data with phenotypic observations will help determine whether genotypes with admixed ancestry ($K=4$) also share the morphological characteristics of both landraces. As already reported in [13], B73 showed very high levels of genetic dissimilarity (>90%) with all the landrace populations.

Rosso Piave, whose production is mainly confined to the extreme east of the Veneto region, showed the lowest mean similarity values in pairwise comparisons with the other three landraces. Consistent with these results, Rosso Piave grouped in a cluster apart from the other varieties for both $K=4$ and $K=6$ (Figure 4).

Although for $K=4$ Biancoperla was represented by a unique group with individual memberships almost always (57 of 62 samples) higher than 80% (Figure 4a), the clustering of individuals for $K=6$ revealed that this variety was totally (100%) constituted by admixed individuals and each individual showed a variable percentage of membership to both clusters (Figure 4b). Since two main accessions of Biancoperla are currently bred in Veneto (ITA0340323 and ITA0340324, which differ in plant size, spike length and color of kernel; [39], according to STRUCTURE results, we speculate that the current Biancoperla population could be the result of repeated events of hybridization and/or introgression between these two accessions.

Conservation of the genetic resources in the agro-ecosystem in which they have evolved (*in situ* conservation) is now being more widely considered as complementary to strategies based on gene banks (*ex situ* conservation, [6]). By taking advantage of the molecular markers and population genetics data here presented, an attempt to increase the genetic purity and to improve the genetic stability of these very old landraces could be made. For each population, only individuals characterized by the highest within-population genetic similarity and ancestry estimates could be maintained and multiplied by open pollination in isolated fields to yield farmer's seed stocks.

Acknowledgments

This work was supported by the project 'PROGRAMMA BIO.NET, rete regionale per la conservazione e caratterizzazione della biodiversità di interesse agrario - Gruppo di lavoro cerealicolo' funded by Programma di Sviluppo Rurale per il Veneto 2007–2013.

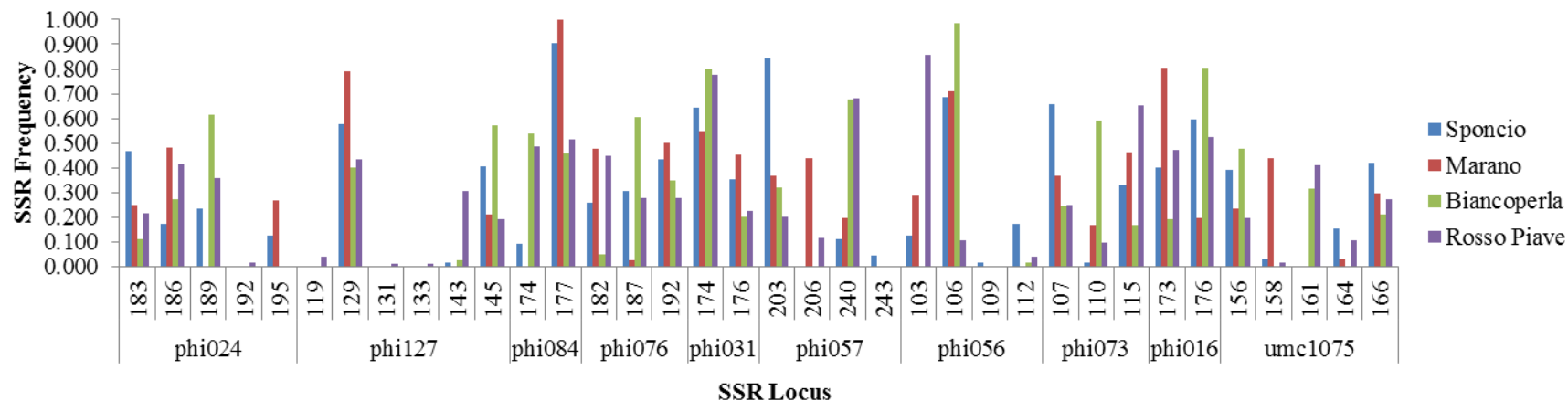
References

1. Baur E. Die Bedeutung der primitiven Kulturrassen und der wilden Verwandten unserer Kulturpflanzen für die Pflanzenzucht. *Jahrb der Dtsch Landwirtschafts Gesellschaft*. 1914;29:104–10.
2. Frankel O, Bennett E. *Genetic Resources in Plants – IBP Handbook No 11*. London: International Biological Programme; 1970.
3. Harlan J. Our vanishing genetic resources. *Science* (80-). 1975;188:618–21.
4. Brush S. In Situ Conservation of Landraces in Centers of Crop Diversity. *Crop Sci*. 1995;35(2):346–54.
5. Bellon MR. Conceptualizing Interventions to Support On-Farm Genetic Resource Conservation. 2004;32(1):159–72.
6. Lucchin M, Barcaccia G, Parrini P. Characterization of a flint maize (*Zea mays* L . convar . mays) Italian landrace : I . Morpho-phenological and agronomic traits. 2001;315–27.
7. Eurostat crop statistics. No Title [Internet]. 2017 [cited 2017 Jul 31]. Available from: <http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do>
8. Shanbao Q, Yuhua W, Tingzhao R, Kecheng Y, Shibin G, Guangtang P. Effective improvement of genetic variation in Maize lines derived from R08xDonor Backcrosses by SSRs. *Biotechnology*. 2009;8(3):358–64.
9. Cömertpay G, Baloch FS, Kilian B, Ülger AC, Özkan H. Diversity Assessment of Turkish Maize Landraces Based on Fluorescent Labelled SSR Markers. *Plant Mol Biol Report*. 2012;30(2):261–74.
10. Pineda-Hidalgo K V., Méndez-Marroquín KP, Alvarez EV, Chávez-Ontiveros J, Sánchez-Peña P, Garzón-Tiznado JA, et al. Microsatellite-based genetic diversity among accessions of maize landraces from sinaloa in méxico. *Hereditas*. 2013;150(4):53–9.
11. Qi-Lun Y, Ping F, Ke-Cheng K, Guang-Tang P. Genetic diversity based on SSR markers in maize (*Zea mays* L.) landraces from Wuling mountain region in China. *J Genet*. 2008;87(3):287–91.
12. Oppong A, Bedoya CA, Ewool MB, Asante MD, Thomson RN, Adu-Dapaah H, et al. Bulk genetic characterization of Ghanaian maize landraces using microsatellite markers. Vol. 59, *Maydica*. 2014. p. 01–8.
13. Barcaccia G, Lucchin M, Parrini P. Characterization of a flint maize (*Zea mays* var. *indurata*) Italian landrace, II. Genetic diversity and relatedness assessed by SSR and Inter-SSR molecular markers. *Genet Resour Crop Evol*. 2003;50(3):253–71.
14. Cooperativa agricola La Fiorita. Mais Sponcio - Cooperativa Agricola La Fiorita [Internet]. 2016 [cited 2016 Dec 31]. Available from: http://www.cooperativala Fiorita.it/?scheda_prodotto+prodotti=mais_sponcio
15. Maresio Bazolle A. *Il possidente bellunese*. Maresio Bazolle A, Perco D, editors. Feltre (Italy): Comunità Montana Feltrina; 1986. 960 p.
16. Zapparoli T. Il granoturco Marano. *L'Italia Agric*. 1939;76:155–9.
17. Zaccaria L. *Valutazione delle variazioni dei caratteri morfo-fisiologici intervenute nel corso degli anni nella varietà di mais Marano Vicentino*. University of Padua; 2012.
18. Agostinetti G. *Cento e dieci ricordi che formano il buon fattor di villa*. Biblioteca dell'Immagine, editor. Pordenone (Italy); 2004. 148 p.
19. Miniscalco V. *Il granoturco*. Cosarini AGF Il., editor. Pordenone (Italy); 1946. 238 p.
20. Clemens R. *Keeping farmers on the land: adding value in agriculture in the Veneto region of Italy*. MATRIC Briefing Papers. 2004.

21. Cassidy A, Mukamal KJ, Liu L, Franz M, Eliassen AH, Rimm EB. High anthocyanin intake is associated with a reduced risk of myocardial infarction in young and middle-aged women. *Circulation*. 2013;127(2):188–96.
22. Schuelke M. An economic method for the fluorescent labeling of PCR fragments A poor man ' s approach to genotyping for research and high-throughput diagnostics . *Nat Biotechnol*. 2000;18:233–4.
23. Register JI, Sullivan H, Yun Y, Cook D, Vaske D. A set of microsatellite markers of general utility in maize. *Maize Genet Coop News Lett*. 2001;75:31–4.
24. Nagy S, Poczai P, Cernák I, Gorji AM, Hegedus G, Taller J. PICcalc: An online program to calculate polymorphic information content for molecular genetic studies. *Biochem Genet*. 2012;50(9–10):670–2.
25. Peakall R, Smouse PE. GenALEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*. 2012;28(19):2537–9.
26. Yeh F, Rong-cai Y, Boyle T, Freeware MW. POPGENE, the user friendly shareware for population genetic analysis. Molecular Biology and Biotechnology Centre, University of Alberta. Alberta, Canada; 1997.
27. Nei M. Analysis of gene diversity in subdivided populations. *Proc Nat Acad Sci*. 1973;70(12):3321–3.
28. van Zonneveld M, Scheldeman X, Escribano P, Viruel MA, van Damme P, Garcia W, et al. Mapping genetic diversity of cherimoya (*Annona cherimola* mill.): Application of spatial analysis for conservation and use of plant genetic resources. *PLoS One*. 2012;7(1).
29. Wright S. The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution (N Y)*. 1965;19(3):395–420.
30. Barcaccia G, Felicetti M, Galla G, Capomaccio S, Cappelli K, Albertini E, et al. Molecular analysis of genetic diversity, population structure and inbreeding level of the Italian Lipizzan horse. *Livest Sci*. 2013;151(2–3):124–33.
31. Rohlf FJ. NTSYS-pc: Numerical Taxonomy and Multivariate Analysis System, version 2.0. Port Jefferson, New York: Applied Biostatistics Inc; 2008. p. 37.
32. Hammer Ø, Harper DAT a. T, Ryan PD. PAST: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontol Electron*. 2001;4(1)(1):1–9.
33. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*. 2003;164(4):1567–87.
34. Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo Á, Lareu M V. An overview of STRUCTURE: Applications, parameter settings, and supporting software. *Front Genet*. 2013;4(MAY):1–13.
35. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol Ecol*. 2005;14(8):2611–20.
36. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*. 1980;32(3):314–31.
37. Palumbo F, Galla G, Barcaccia G. Developing a Molecular Identification Assay of Old Landraces for the Genetic Authentication of Typical Agro-Food Products : The Case Study of the Barley “ Agordino .” *Food Technol Biotechnol*. 2017;55(1):29–39.
38. Russell WA. Genetic Improvement of Maize Yields. *Adv Agron*. 1991;46:245–98.
39. Rete regionale per la conservazione delle varietà tradizionali appartenenti alle principali specie agrarie del Veneto. *Biodiversità del Veneto*. 2005.

Supplementary materials

Figure S1. Allele frequencies by population



Locus	Allele	Sponcio	Marano	Biancoperla	Rosso Piave	Locus	Allele	Sponcio	Marano	Biancoperla	Rosso Piave
phi024	183	0.469	0.250	0.113	0.214	phi057	203	0.844	0.367	0.323	0.200
	186	0.172	0.484	0.274	0.414		206	0.000	0.438	0.000	0.117
	189	0.234	0.000	0.613	0.357		240	0.109	0.195	0.677	0.683
	192	0.000	0.000	0.000	0.014		243	0.047	0.000	0.000	0.000
	195	0.125	0.266	0.000	0.000		phi056	103	0.125	0.289	0.000
phi127	119	0.000	0.000	0.000	0.038	106		0.688	0.711	0.984	0.105
	129	0.578	0.789	0.403	0.436	109		0.016	0.000	0.000	0.000
	131	0.000	0.000	0.000	0.013	112		0.172	0.000	0.016	0.039
	133	0.000	0.000	0.000	0.013	phi073	107	0.656	0.366	0.242	0.250
	143	0.016	0.000	0.024	0.308		110	0.016	0.170	0.589	0.097
phi084	145	0.406	0.211	0.573	0.192	115	0.328	0.464	0.169	0.653	
	174	0.094	0.000	0.540	0.487	phi016	173	0.403	0.805	0.194	0.474
177	0.906	1.000	0.460	0.513	176		0.597	0.195	0.806	0.526	
phi076	182	0.258	0.476	0.048	0.447	umc1075	156	0.391	0.234	0.476	0.197
	187	0.306	0.024	0.605	0.276		158	0.031	0.438	0.000	0.015
	192	0.435	0.500	0.347	0.276		161	0.000	0.000	0.315	0.409
phi031	174	0.645	0.547	0.798	0.776	164	0.156	0.031	0.000	0.106	
	176	0.355	0.453	0.202	0.224	166	0.422	0.297	0.210	0.273	

Figure S2. Mean genetic similarity matrix considering all pairwise comparisons

	Sponcio	Marano	Biancoperla	Rosso Piave
Sponcio	74.71 (± 9.80)			
Marano	67.23 (± 7.93)	75.40 (± 8.27)		
Biancoperla	65.50 (± 7.79)	59.10 (± 7.14)	78.12 (± 7.60)	
Rosso Piave	60.00 (± 8.15)	60.50 (± 8.25)	62.80 (± 7.29)	69.98 (± 10.55)

Chapter IV

The leaf transcriptome of fennel enables the characterization of the t-anethole pathway and the discovery of microsatellites and single-nucleotide variants

Abstract

Despite its agronomic and pharmaceutical interest, fennel is a vegetable species characterized by genetic and molecular data shortage. Taking advantage of NGS technology, we sequenced and annotated the first fennel leaf transcriptome using four different varietal genotypes and following two different bioinformatics approaches: *de novo* and genome-guided transcriptome assembly. A reference transcriptome was produced combining these two types of transcriptomes. Amongst the 79,263 loci obtained, 47,775 were annotated by mean of BLASTx analysis performed against a NR protein database subset, with 11,853 loci representing putative full-length CDS. Bioinformatics analyses revealed 1,011 transcripts encoding for transcription factors, mainly from the BHLH, MYB-related, C2H2, MYB, and ERF families and 6,411 EST-SSR regions. Single-nucleotide variants including SNPs and In/Dels were identified among the four genotypes, with a frequency of 0.5 and 0.04 variants per Kb, respectively. Finally, assembled transcripts were screened for identifying genes related to the phenylpropanoid pathway and, in particular, to the biosynthesis of t-anethole, a compound well-known for its nutraceutical and medical properties. Nine transcripts significantly matched with anethol-related biosynthetic genes. Overall, our work represents a treasure trove of information exploitable both for marker-assisted breeding and for in-depth studies on thousands of genes, including those involved in t-anethole biosynthesis.

Keywords: EST-SSR, SNP, phenylpropanoid, de novo assembly, genome-guided assembly, NGS

Introduction

Foeniculum vulgare Mill. ($2n=2x=22$), commonly known as fennel, is a biennial or perennial diploid species belonging to Apiaceae family (or Umbelliferae, *nomen conservandum*). It originated in the southern Mediterranean region and, through naturalization and cultivation, it widely spread all over the world, specifically on dry soils near the coastal areas and river banks in Asia, North America and Europe. The wild as well as cultivated forms of fennel are hermaphrodite: the species of agronomic and pharmaceutical interest reproduces prevalently by outcrossing, but selfing is also possible. Most cultivated varieties are OP synthetics, although F1 hybrids have been bred in recent years. Fennel is cultivated both for its inflated leaf bases, which form an edible bulb-like structure, eaten as a raw or cooked vegetable, and for seeds, appreciated for their pleasant fragrance and aromatic taste. FAO statistics highlight the economic impact of this species, revealing that India is the world leader producer of fennel with more than 500,000 tons per year, followed by Mexico and China [1].

Fennel is a species of great interest also because of its pharmaceutical properties, since it accumulates several compounds with beneficial effects on human health. At this regard, t-anethole is the major component of the essential oils produced by leaves, with a reported content reaching up to 97.1% of the total volatile compounds, and with concentrations that vary considerably depending on the phenological state and geographical origin [2]. This organic compound has been extensively explored standing out for its capability in reducing mild spasmodic gastro-intestinal pains [3] as well as for its antithrombotic [4] and hypotensive [5] activities. For what concerns the biological role of anethole *in planta*, this specific compound belongs to the group of phytoalexins, acting as antimicrobial [6], antifungal [7] and insecticidal [8], being generally related to plant defense from biotic stresses.

Despite its agronomic and pharmaceutical interest, molecular data available for fennel are scanty and, to the best of our knowledge, only few genetic studies have been performed on this species. As

a matter of fact, most of the genic and genomic DNA sequences currently available on public databases (*e.g.* GenBank) concern the chloroplast genome, whose draft sequence was published, together with those of dill (*Anethum graveolens*) and coriander (*Coriandrum sativum*), to highlight the extent of large inverted repeat variation among some taxa of the Apiaceae family [9].

Until a few years ago, most of the molecular information available to elucidate various complex biological phenomena was deriving from extensive investigations on few model plants. Nowadays, the recent advances in Next-Generation Sequencing (NGS) technologies and the sticking reduction of DNA sequencing costs led to a raise in the number of transcriptomic studies also in non-model plant species [10]. Although the major goal of RNA-seq analyses relies on the identification of differentially expressed genes (DEGs) amongst different conditions, organs or tissues and developmental stages, this technology is also very useful for the identification of expressed transcripts (ESTs) related to genes involved in metabolic pathways of interest and for the detection of genetic variations such as single nucleotide polymorphisms (SNPs) and simple sequence repeats (SSRs).

In this study, we took advantage of NGS technology to perform the first fennel leaf transcriptome sequencing and *in silico* assembly by using two different approaches: *de novo* without the aid of a sequenced genome and genome-guided transcriptome assemblies. The first strategy, particularly useful for those organisms without a reference genome, is based on the reconstruction of contigs by overlapping the reads obtained from the sequencer, taking advantage of their high level of redundancy. For those species with a reference genome available, a genome-guided assembly is generally preferable. This second strategy aligns reads to a reference genome to finally assemble overlapping alignments into transcripts [11].

This paper deals with a comparative analysis of the transcriptome assembly strategies with the aim of understanding their main features and differences. The newly assembled leaf transcriptome dataset will be presented and critically discussed, along with its utility as an important resource for further genetic characterization of cultivated fennel accessions. Specific emphasis will be given to

the characterization of the main genes and gene products involved in the t-anethole biosynthetic pathway and the identification of single-nucleotide variants exploitable in advanced breeding programs in *F. vulgare*.

Materials and Methods

DNA/RNA isolation and sequencing

Transcriptome sequencing (RNA-seq) of the leaves was conducted using four agronomically relevant breeding lines of cultivated fennel, namely OL1, OL2, OL9 and OL164, in two biological replicates, each one constituted of single individuals from each genotype. Plant materials used in this study were chosen based on the following criteria: i) commercial importance of the variety to which each line belongs; ii) robust phenotypic and genotypic characterization available for each breeding line; iii) high degree of homozygosity (> 90%); iv) representativeness of four cultivated biotypes showing distinct esthetical, agronomic and aromatic traits as well as unrelated genetic backgrounds.

The OL2 genotype, being characterized by the highest degree of homozygosity (93%) was also employed for genome sequencing.

For each line, seeds were sown and plants grown under standard cultivation conditions. Leaves were collected from 1-month-old individuals, snap-frozen in liquid nitrogen upon harvesting and stored at -80°C until further processing. Total RNA was extracted using RNeasy Plant Mini Kit (Qiagen GmbH, Hilden, Germany), and treated with RNase-Free DNase set (Qiagen GmbH, Hilden, Germany) according to manufacturer's instructions.

Genomic DNA was extracted from leaf tissues using a standard CTAB protocol [12]. The quality of nucleic acids was estimated by spectrophotometric analysis (NanoDrop 2000c UV-Vis, Thermo Fisher Scientific) and agarose gel electrophoresis (1.0% w/v agarose TAE 1× gel containing 1× SYBR® Safe, Thermo Fisher Scientific). In addition, the integrity of RNA samples was analyzed using the RNA 6000 Pico Kit (Agilent Technologies, Santa Clara, CA) on a Bioanalyzer 2100

(Agilent Technologies). Samples with RIN (RNA Integrity Number) values of at least 7 (*i.e.* $RIN \geq 7$) were considered suitable for the following steps.

An equal amount of total RNA (1 μ g) from one individual plant of each fennel accession was used as input for the TruSeq Stranded mRNA LT Sample Prep Kit (Illumina, Inc., San Diego, CA, USA) and, by means of indexed adapters, a sequencing library was created according to the manufacturer's instructions. The library was then sequenced on a HiSeq 2000 instrument (Illumina, Inc., San Diego, CA, USA) with paired-end, 100-bp-read chemistry. A total of 2 μ g of genomic DNA were subjected to library preparation for whole-genome sequencing using the Illumina TruSeq DNA PCR-free sample preparation kit (Illumina, Inc., San Diego, CA, USA) according to the instructions provided by the company. The library was sequenced on an Illumina HiSeq 2500 using paired-end, 150-bp-read chemistry (Illumina, Inc., San Diego, CA, USA).

Genome draft and leaf transcriptome assembly

Raw genomic sequences were processed with Trimmomatic software [13] to remove the adapter sequences and to trim low quality bases. In particular, Trimmomatic was run setting an average minimum quality score of 20 within a sliding window of 5 and the minimum reads length was set to 75 bp. The filtered sequences were assembled using SOAPdenovo2 [14] into contigs at distinct k-mer values (from 71 to 121). Sequence statistics were calculated using a Perl script, NGSQCToolkit_v2.3.3 [15]. In order to assess the quality of the assembly and to estimate and to validate the portion of the assembled coding sequence, we aligned the RNA-seq reads against our newly constituted genome draft of fennel.

Raw RNA sequence data were filtered using standard RNA-Seq parameters by means of CLC Genomics Workbench 7.0.4 (CLCbio, Aarhus, Denmark). Briefly, raw reads were demultiplexed and the 3' ends were trimmed to form eight sets of reads from the four different varietal genotypes (each one in two biological replicate). Reads were then processed as follows for: i) removing low quality sequences with a 0.05 error probability limit; ii) discarding reads with final length < 25 bp;

iii) trimming reads with more than two ambiguous nucleotides. For a *de novo* transcriptome assembly, filtered data coming from the two biological replicates of each variety were assembled using CLC Genomics Workbench 7.0.4 (CLCbio, Aarhus, Denmark) run at default settings. The four transcriptome assemblies obtained, one for each varietal genotype/line, were merged into a global one, considered as a *de novo* leaf reference transcriptome. For the *ex post* genome-guided assembly, leaf transcriptome data obtained from the two biological replicates of each variety were filtered and then aligned against the draft reference genome newly developed using HISAT [16] at default settings. StringTie [17] was then used to reconstruct the transcriptome of each variety and to collapse them into a global genome-guided leaf reference transcriptome, using the “merge” option. A final clustering was accomplished overlapping the two newly assembled *de novo* and genome-guided transcriptomes: CD-HIT [18] was used to cluster all sequences with a similarity threshold > 95% and to generate a third transcriptome designed as “cluster transcriptome”. The quality of the three transcriptome assemblies (*de novo*, genome-guided and cluster transcriptomes) was then assessed using NGSQCToolkit_v2.3.3. Total number of reads/sequences in the file, total and individual (A, T, C, G and N) number of bases, G+C and A+T counts, and minimum, maximum, average, median, N25, N50, N75, N90 and N95 length N50 value, were evaluated.

Functional annotation and classification

Putative transcripts scaffolds of the fennel clustered transcriptome were validated comparing them with a subset of the NR protein database focused on the Pentapetalae clade, using a BLASTx-based approach (E-value $\leq 1e^{-05}$, BLAST v.2.3.0+). Assumed that each assembled locus represented a single gene, the best hit for each transcript was selected. Moreover, in order to extrapolate Gene Ontology annotations [19] and KEGG terms [20], the GI identifiers of the BLASTx hits were mapped to the UniprotKB protein database [21]. In addition, a locus was predicted as full-length transcript if the ratio between its BLASTx alignment length and the subject length extrapolated from UniprotKB protein database was higher than 0.95.

Finally, further enrichment of enzyme annotations was made with the BLAST2GO software v1.3.3 [22] using the function “direct GO to Enzyme annotation” to perform basic statistics on ontological annotations, reducing the complexity of the data.

RNA-Seq data were also used to conduct the identification of loci related to the phenylpropanoid pathway and to the t-anethole/methylchavicol biosynthesis. All the amino acid sequences available in NCBI and associated to the enzymes involved in these two pathways were retrieved and aligned based on a tBLASTn-based approach (E-value $\leq 1e^{-20}$) against the cluster transcriptome, used as nucleotide database. The best hit for each enzyme was selected and, through BLASTn analysis (E-value $\leq 1e^{-40}$), aligned against the *D. carota* transcriptome [23], which belongs to the same family (Apiaceae). PlantTFDB [24] was used to translate scaffolds from the cluster transcriptome to protein and to predict *in silico* transcription factors (TF, E-value $\leq 1e^{-05}$). Each prediction was further linked to the best hit in *Arabidopsis thaliana*. The results were finally compared with the TF abundance in the transcriptome of *D. carota*, from the same Apiaceae family.

Simple sequence repeats (SSRs) identification

Simple sequence repeat regions were detected using the MicroSatellite (MISA) Identification Tool Perl script [25]. The assembled sequences were screened for di-, tri-, tetra-, penta- and hexa-nucleotide repeat motifs with a minimum repeat number of 7, 6, 6, 6 and 5, respectively. The maximal number of nucleotides interrupting two SSR regions in a compound microsatellite was set at 100 bp and the space between imperfect SSR stretches was set at 5 bp.

Single nucleotide polymorphisms (SNPs) identification

The raw reads were processed for adapter removal, quality trimming and filtering for organelle DNA and duplicates. Post-processed paired-end reads longer than 50 bp were aligned to the cluster reference transcriptome using BWA [26] with default parameters. Local realignment around In/Dels was performed with the RealignerTargetCreator and IndelRealigner tools of the GATK package, version 2.1-13 [27]. Variant positions were identified using the HaplotypeCaller tool of the GATK

package with default parameters. Depth-of-coverage was analyzed using DNACopy. SNP variant were hard filtered using the VariantFiltration tool according to GATK instruction. The pairwise genetic distance among the 8 genotypes (4 breeding lines in two biological replicates) was estimated based on Nei's [28] unbiased genetic distance using R tools vcfR and Adegenet.

Results

Genome draft and transcriptome assembly results

Among the four varietal genotypes of cultivated fennel selected within this work, the highly homozygous OL2 accession was chosen for genome sequencing. A total of 486,073,396 paired end reads, corresponding to 72.91 Gbp, were generated by means of an Illumina HiSeq 2500 platform. The assembly resulted highly fragmented and split into several small contigs (7,978,334 contigs, N50=319). Considering the assembled scaffolds, we obtained a total of 300,408 sequences whose length ranged from 370 to 145,787 bp, for a total of 1.01 Gbp (N50=9,443).

Table 1 reports the main descriptive statistics. A total number of 41,192,930,800 bp, corresponding to 407,850,800 paired-end 101 bp raw reads, was obtained from Illumina mRNA sequencing of *F. vulgare* leaves. Considering the biological replicates for each genotype, on average $101,962,700 \pm 9,372,035$ reads were produced. Raw data were trimmed removing both the Truseq Universal and Indexed adapters and, after removal of low quality sequences, 392,798,001 reads with final length higher than 25 bp were used for the assembly. Of these, 83.74% reached an average PHRED score threshold of $Q \geq 35$.

The leaf transcriptome coverage resulted very high since 91% of the RNA-seq reads covered correctly and were properly aligned with the newly developed genome draft.

Using two different approaches (*de novo* and genome-guided), the reads were firstly assembled into four distinct transcriptomes, one for each varietal genotype, adopting optimal parameters. Therefore, the four transcriptomes obtained from each strategy were further merged to produce two different leaf reference transcriptomes: a *de novo* transcriptome and a genome-guided

transcriptome. In this last step, the CLC Genomics Workbench platform enabled to assemble 61,299 transcripts, whereas the StringTie assembler generated a total of 51,917 transcripts. A clustering of these two assemblies, performed using CD HIT, was used for subsequent analysis since the number of assembled loci and total bases was higher compared to the other two assembled transcriptomes considered singularly (Table 1). In detail, the “clustered transcriptome” contained up to 79,263 assembled loci with an average length of 1,142 bp, N50 length of 1,654 bp and maximal length of 14,975. Overall, 90,513,363 bp were assembled.

Raw transcriptome sequences files are available on the Sequence Read Archive (SRA) with the following accession numbers: SSR6265712-SSR6265719. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GGAC00000000. The version described in this paper is the first version, GGAC01000000.

Table 1. Main descriptive statistics related to fennel genome and transcriptome assemblies. Statistics are available for the genome draft, *de novo* transcriptome assembly, the genome-guided transcriptome assembly and the clustered transcriptome assembly

Main Statistics	Genome	Transcriptome		
		<i>De novo</i>	Genome-guided	Clustering
Total sequences	300,408	61,299	51,917	79,263
Total bases	1,011,093,015	57,419,229	68,310,969	90,513,363
Min sequence length	370	104	200	115
Max sequence length	145,787	14,975	13,065	14,975
Average sequence length	3365.73	936.71	1,315.77	1,141.94
Median sequence length	1241.00	627	1,073	824
N25 length	20734	2,313	2,681	2,545
N50 length	9443	1,353	1,850	1,654
N75 length	2842	705	1,162	936
N90 length	1126	425	684	535
N95 length	777	333	471	396
As %	31.81	31.39	29.61	30.62
Ts %	31.72	29.68	30.13	30.06
Gs %	16.06	18.31	21.28	19.76
Cs %	16.11	20.61	18.62	19.3
(A+T)s %	63.53	61.08	59.74	60.68
(G+C)s %	32.17	38.92	39.89	39.06
Ns %	4.30	0	0.37	0.26

Functional classification of the clustered transcriptome

BLASTx analysis (E-value $\leq 1e^{-05}$) performed against the Pentapetalae clade subset of the NR protein database, identified up to 47,775 transcripts (60.27% of the total transcript number) showing a significant match and, amongst these, 8,067 (16.88%) and 4,868 transcripts (10.19%) were related

to sequences from *Vitis vinifera* and *Sesamun indicum*, respectively (see Supplementary Figure S1). Considering the similarity and E-value distribution, 10,152 (21.25%) assembled sequences showed similarity scores higher than 80% and 19,850 (41.55%) loci exhibited extremely low E-values ($\leq 1e^{-100}$, Supplementary Figures S2 and S3). Overall, a total of 11,853 (24.81%) loci contained a putative full-length CDS and among them 2,392 (5.01%) revealed similarity scores higher than 80% with their best hit subject. The 47,775 fennel transcripts showing a BLASTx match were imported in Uniprot for GO mapping and EC annotation. 36,204 GO IDs and 1,629 EC number were assigned respectively to 14,734 and 1,615 fennel transcripts. 11,799 GO IDs (32.59%) were assigned to the “biological process” category (BP), 14,771 (40.80%) to the “cellular component” category (CC), and 9,634 (26.61%) to the “molecular function” category (MF). The GO IDs were distributed in 15 levels among these three categories and based on highest number of annotated GO terms, the most informative GO level resulted to be level 5, retrieving 7,738 GO IDs (Figure 1).

Therefore, level 5 was used to summarize the GO terms in subcategories (Figure 2). In the MF category “nucleotide binding” (13%), “ribonucleoside binding”, “purine nucleoside binding” and “metal ion binding” (10%, each) ontologies were more abundant. Moreover, “intracellular part” (48%) and “integral to membrane” (42%) represented almost the totality of the CC categories whereas “cellular macromolecule metabolic process” (8%), “nucleobase-containing compound metabolic process” and “gene expression” (5%, each) and “macromolecule biosynthetic process” and “protein metabolic process” (both 4%) were the dominant subcategories inside the BP category.

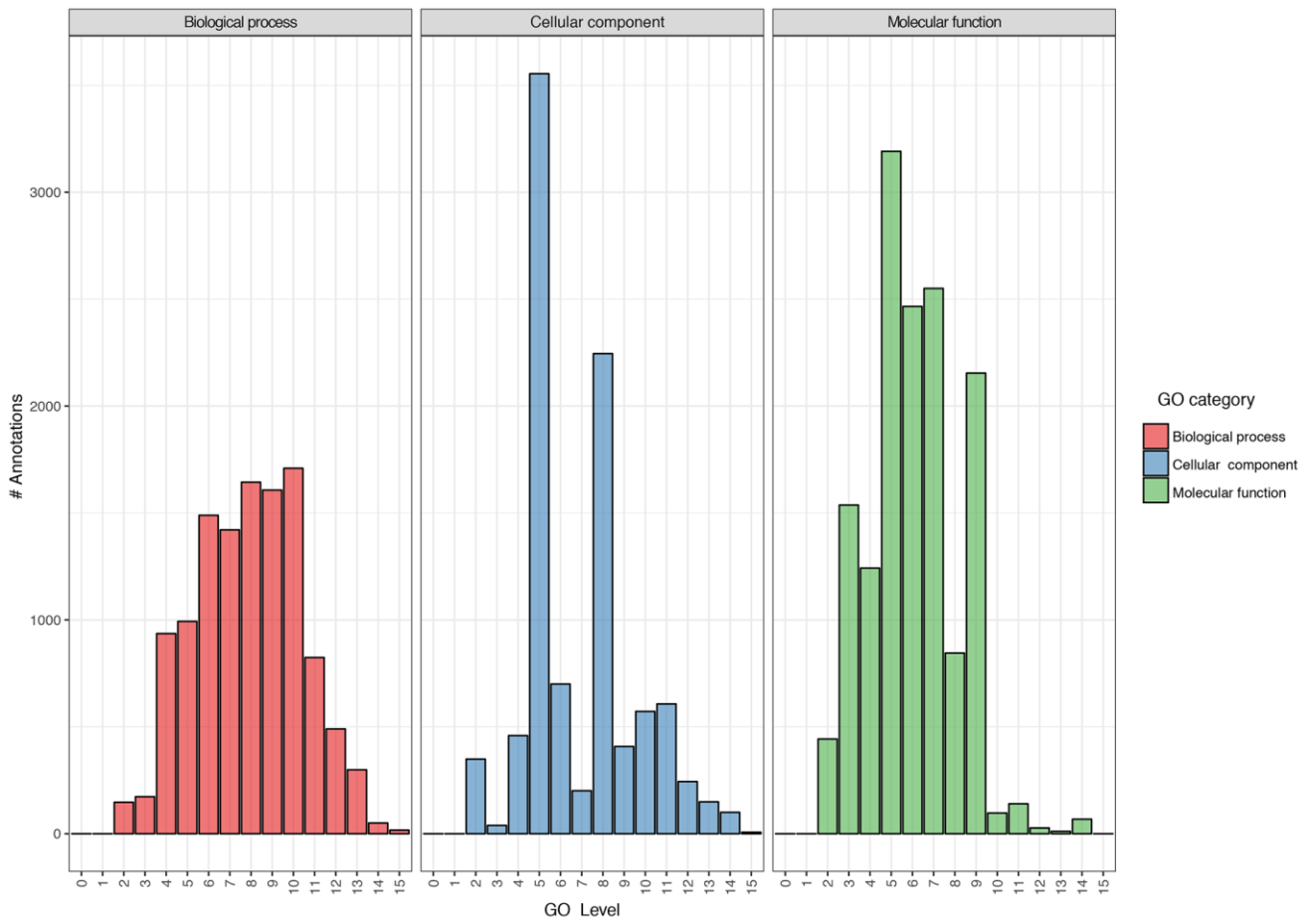


Figure 1. Gene ontology (GO) level distribution chart for the fennel leaf transcriptome according to the Biological Process (BP), Molecular Function (MF) Cellular Component (CC) categories

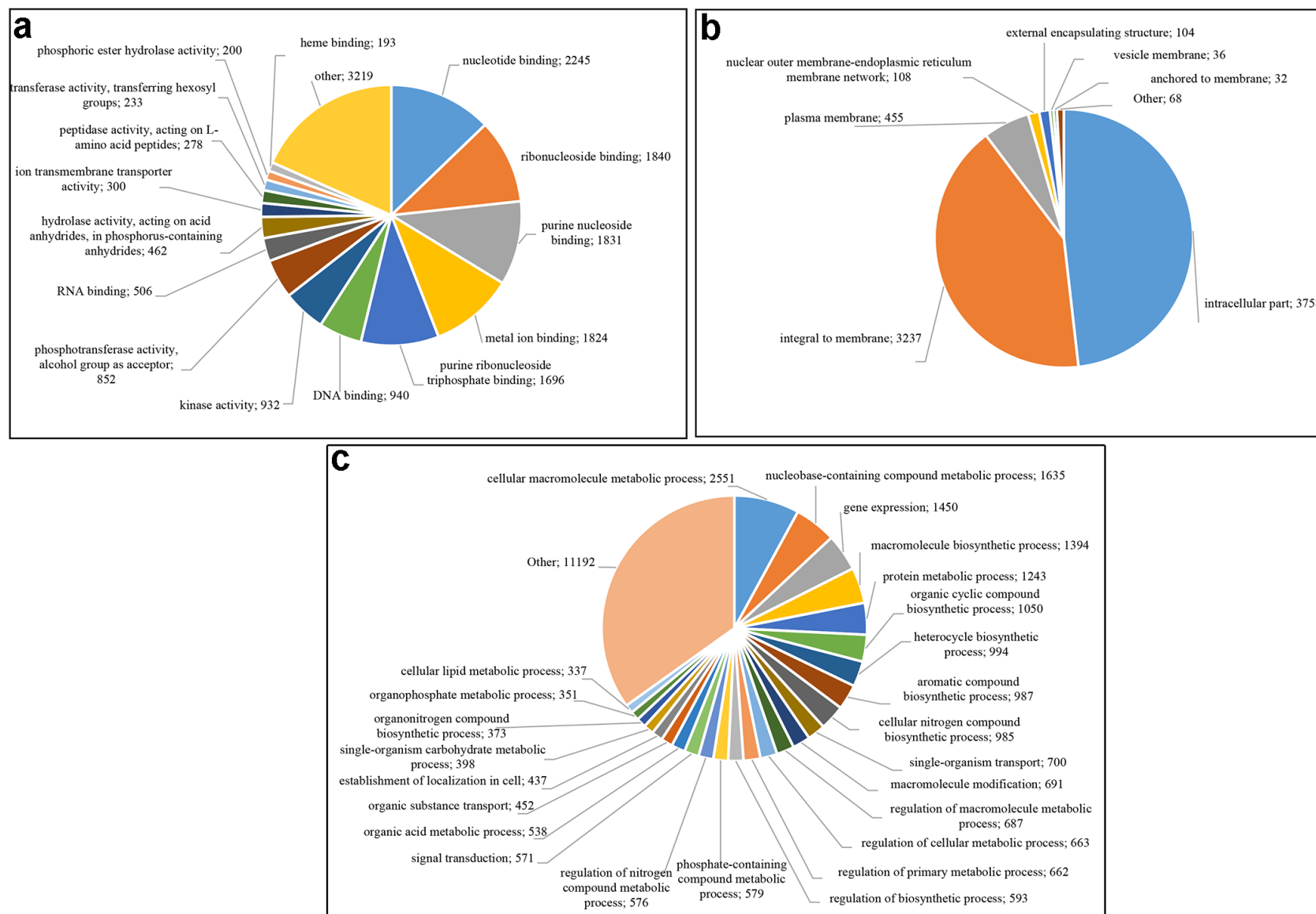


Figure 2. Gene ontology (GO) classification of assembled loci. The results of BLASTX searches against the Pentapetalae clade subset of the NR protein database were used for GO term mapping and annotation. The number of sequences assigned to level 5 GO terms for GO subcategories including molecular function (a), cellular component (b), and biological process (c), are shown

Transcription factor identification

Amongst the 79,263 contigs reconstructed in the clustered transcriptome, we were able to identify 1,011 leaf transcripts encoding for transcription factors (TFs) based on the plant transcription Factor database (PlantTFDB). The abundance of each different multigenic family was evaluated in *F. vulgare* and in *D. carota* (Figure 3). Amongst them, BHLH, MYB-related, C2H2, MYB, ERF and NAC were the six most represented categories in both species.

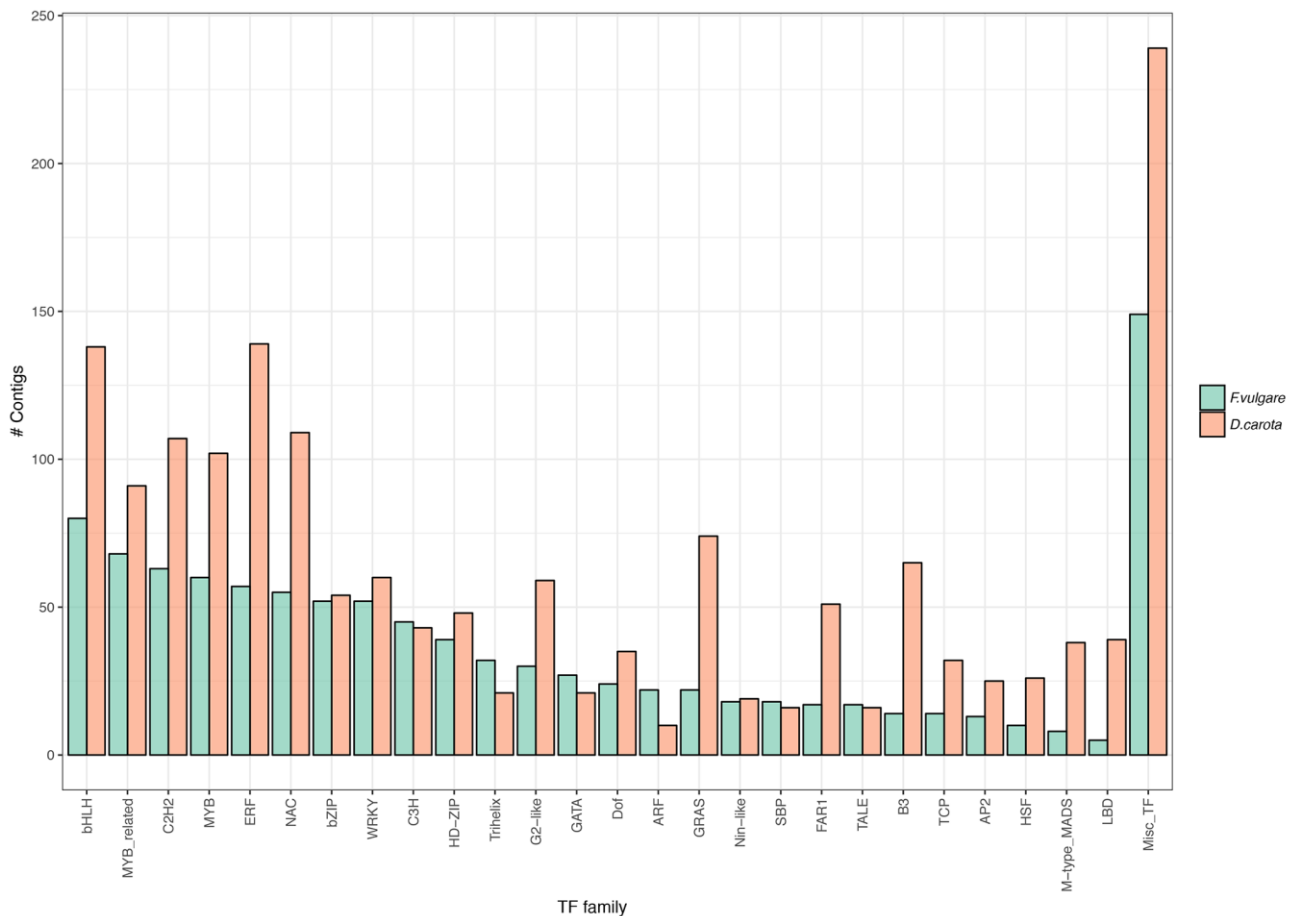
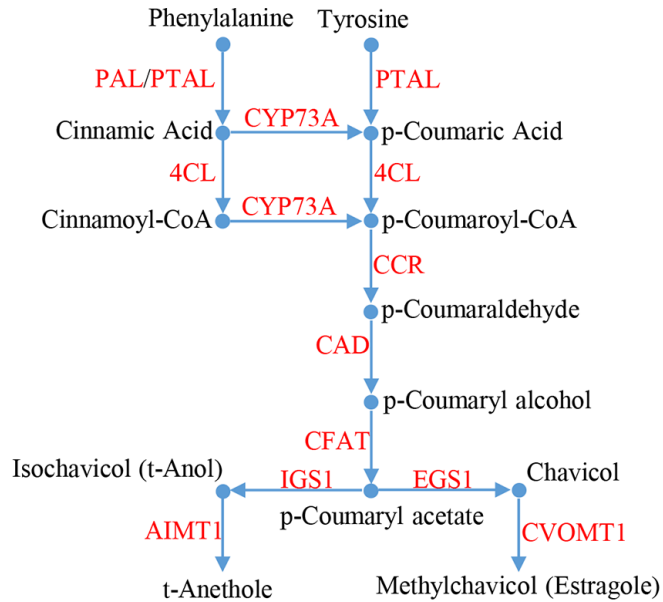


Figure 3. Transcription factor family analysis. Number of transcription factors determined within the fennel combined leaf transcriptome assembly grouped by transcription factor family

Identification of gene transcripts involved in the anethole biosynthetic pathway

For each one of the 11 protein sequences retrieved from NCBI and involved in the biosynthesis of t-anethole/methylchavicol, we identified the fennel transcript showing the most significant match (tBLASTn, E-value $\leq 6e^{-46}$). These 11 candidate transcripts were, in turn, successfully aligned against *D. carota* transcriptome (BLASTn, E-value $\leq 1e^{-122}$) and the NR database (BLASTx, E-

value $\leq 3e^{-64}$). For 7 out of the 11 enzymes involved in the t-anethole/methylchavicol biosynthesis (*i.e.* PAL, PTAL, CYP73A, 4CL, CCR, CAD, CFAT) results obtained with tBLASTn alignment were coherent also with BLASTn and BLASTx (Figure 4). Concerning the t-Anol/isochavicol O-methyltransferase (AIMT1, 2.1.1.279) only the BLASTn alignment of MSTRG.32111 contig against *D. carota* transcriptome confirmed results obtained with tBLASTn. Finally, MSTRG.27089 significantly matched (E-value $\leq 6e^{-122}$) with two different enzymes: chavicol synthase (1.1.1.318) and t-anol/isochavicol synthase (1.1.1.319).



Name	EC Number	Protein
PAL	4.3.1.24	Phenylalanine ammonia-lyase
PTAL	4.3.1.25	Phenylalanine/tyrosine ammonia-lyase
CYP73A	1.14.13.11	Trans-cinnamate monooxygenase
4CL	6.2.1.12	Coumarate-CoA ligase/coumarate-CoA synthase
CCR	1.2.1.44	Cinnamoyl-CoA reductase
CAD	1.1.1.195	Cinnamyl alcohol dehydrogenase
CFAT	2.3.1.-	Coniferyl alcohol acyltransferase
EGS1	1.1.1.318	Chavicol synthase
IGS1	1.1.1.319	t-Anol/isochavicol synthase
CVOMT1	2.1.1.146	Chavicol O-methyltransferase
AIMT1	2.1.1.279	t-Anol/isochavicol O-methyltransferase

t-anethole pathway enzymes			TblastN (t-anethole enzymes vs <i>F.vulgare</i> transcripts)			blastX (TblastN results vs Penatpetalae)					blastN (TblastN results vs <i>D. carota</i> transcripts)			
EC number	gi ref	Source species	Best hit	pident	e-value	gi ref	EC number	Species	pident	e-value	Contig	EC	pident	e-value
4.3.1.24	497421	<i>Arabidopsis thaliana</i>	MSTRG.4098.1	84.10	0	225454653	4.3.1.24	<i>Vitis vinifera</i>	85.99	0	Dck075570	4.3.1.24	91.00	0
4.3.1.25	821595499	<i>Zea mays</i>	FV 1883	70.50	0	747093693	4.3.1.25	<i>Sesamum indicum</i>	88.57	0	Dck017457	4.3.1.25	92.00	0
1.14.13.11	3915085	<i>Arabidopsis thaliana</i>	MSTRG.10629.1	84.33	0	590591697	1.14.13.11	<i>Theobroma cacao</i>	88.47	0	Dck024733	1.14.13.11	90.00	0
6.2.1.12	12229649	<i>Arabidopsis thaliana</i>	MSTRG.14069.1	58.88	0	698507025	6.2.1.12	<i>Nicotiana tabacum</i>	83.95	0	Dck003566	6.2.1.12	89.00	0
1.2.1.44	332191267	<i>Arabidopsis thaliana</i>	MSTRG.19636.1	72.90	9E-160	568880935	1.2.1.44	<i>Citrus sinensis</i>	86.88	0	Dck032472	1.2.1.44	89.00	0
1.1.1.195	15235757	<i>Arabidopsis thaliana</i>	FV 5138	70.93	4E-171	225426492	1.1.1.195	<i>Vitis vinifera</i>	78.04	7E-174	Dck034486	1.1.1.195	92.00	0
2.3.1.-	110559372	<i>Petunia x Hybrida</i>	MSTRG.22730.1	54.84	2E-153	747092508	2.3.1.-	<i>Sesamum indicum</i>	54.71	7E-144	Dck002566	2.3.1.133	84.00	0
1.1.1.318	1052489089	<i>Ocimum basilicum</i>	MSTRG.27089.1	56.37	6E-122	224095730	-	<i>Populus trichocarpa</i>	62.10	6E-140	Dck027129	1.1.1.319	88.00	0
1.1.1.319	218963652	<i>Pimpinella anisum</i>	MSTRG.27089.1	70.50	3E-165	224095730	-	<i>Populus trichocarpa</i>	62.10	6E-140	Dck027129	1.1.1.319	88.00	0
2.1.1.146	16903138	<i>Ocimum basilicum</i>	MSTRG.28896.1	37.28	6E-46	590683948	2.1.1.128*	<i>Theobroma cacao</i>	68.21	3E-64	Dck039291	2.1.1.128*	90.00	E-124
2.1.1.279	218963654	<i>Pimpinella anisum</i>	MSTRG.32111.1	75.00	0	590640752	2.1.1.68**	<i>Theobroma cacao</i>	44.79	1E-98	Dck032876	2.1.1.279	80.00	E-122

*2.1.1.128 RS-norcochlorine 6-O-methyltransferase; **2.1.1.68 Caffeic acid 3-O-methyltransferase 1

Figure 4. Reconstruction of the t-anethole biosynthetic pathway based on the annotation of the fennel combined transcriptome assembly. All enzymes involved in the pathway are schematically reported together with their Enzyme Classification (EC) and abbreviated name. The main BLAST results are summarized in the lower table, including: i) a first tBLASTn approach used to find the best match between the fennel transcriptome and each enzyme of the t-anethole pathway ii) a BLASTx approach used to find the best match between each result of the tBLASTn approach and the Penatpetalae database iii) a BLASTn approach used to find the best match between each result of the tBLASTn approach and the *D. carota* transcriptome

Identification of expressed EST-SSRs and SNPs

The MISA program allowed the identification of 6,411 SSRs in 5,139 transcripts, with 954 sequences containing more than one microsatellite region. Amongst the SSRs, 4,623 were “perfect”, 9 “imperfect” and 1,779 “compound”. The majority (97.6%) of SSRs were detected in the di- and tri-nucleotide categories (82.7 % and 14.6 %, respectively), followed by tetra- (1.2%), esa- (1%), penta-nucleotide categories (0.1%). Results of the EST-SSRs are summarized in Figure 5. The most common type of dinucleotide was AG/CT (64.9%), followed by AC/GT (13.7%). The longest perfect di-nucleotide SSR [(CT)₄₄] was found within the MSTRG.29622 contig, which encodes for a geranylgeranyl transferase component A1 (BLASTx E-value=3e⁻⁶⁰), (*D. carota*), while the longest tri-nucleotide microsatellite [(AAG)₃₂] was found within the MSTRG.22085 contig, matching with the homeobox protein 12 (*D. carota*) (BLASTx E-value=3e⁻⁵²).

A total of 43,237 SNPs and 3,955 In/Dels were also identified among four genotypes analyzed in two biological replicates, applying a SNP calling approach based on deep multiple alignment (minimum 8× coverage) and allowing no more than two missing data. The global inter-samples SNP frequency was 0.5 per Kb and the In/Del frequency 0.04 per Kb. The transition / transversion ratio was 1.65.

The average genetic distance between individual genotypes from the same breeding line ranged from 0.04 within the OL2 line to 0.30 within the OL164 line, whereas the genetic distance between individual genotypes belonging to different lines varied from 0.35 (OL1 vs. OL164) to 0.57 (OL9 vs. OL2; Figure 6).

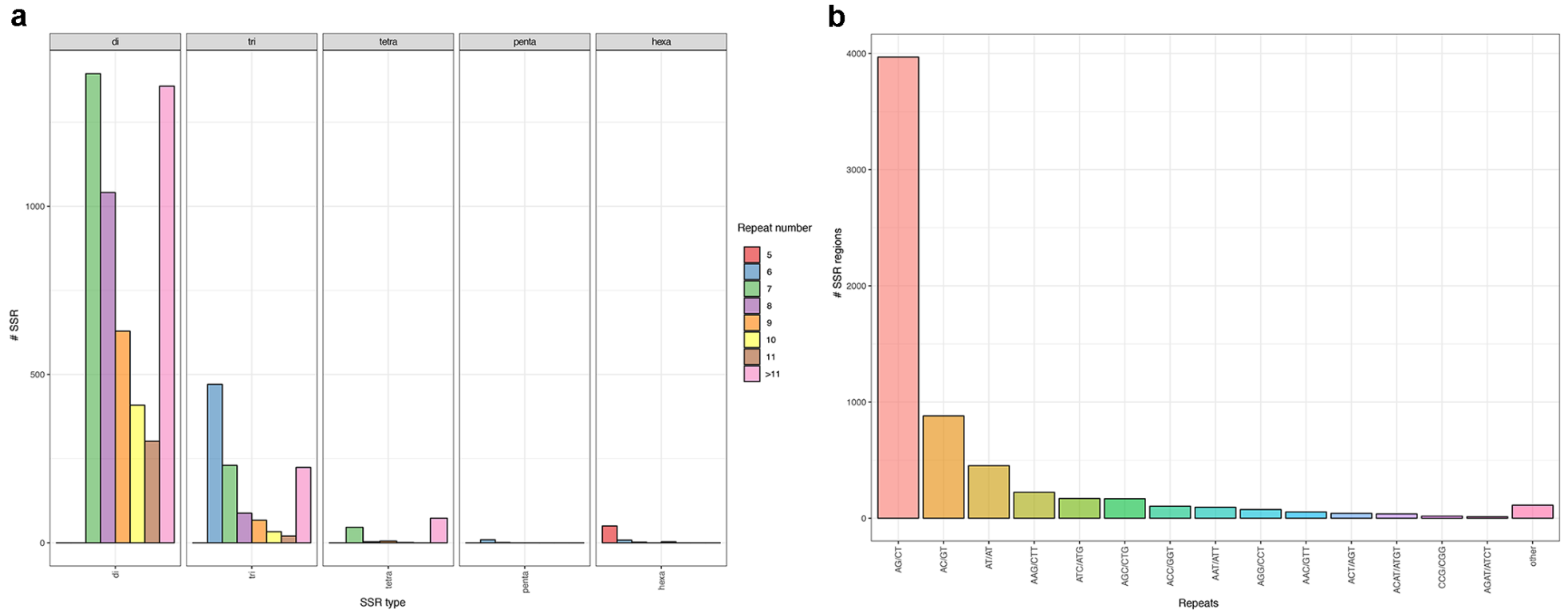


Figure 5. Summary statistics related to EST-SSR regions: a) distribution of the EST-SSR motif repeat numbers from di- to hexa-nucleotide types (the vertical axis shows the abundance of microsatellites with different motif repeat numbers, from 5 up to >11, which are discriminated by different colors as reported in the legend); b) most common types of EST-SSR motif repeats

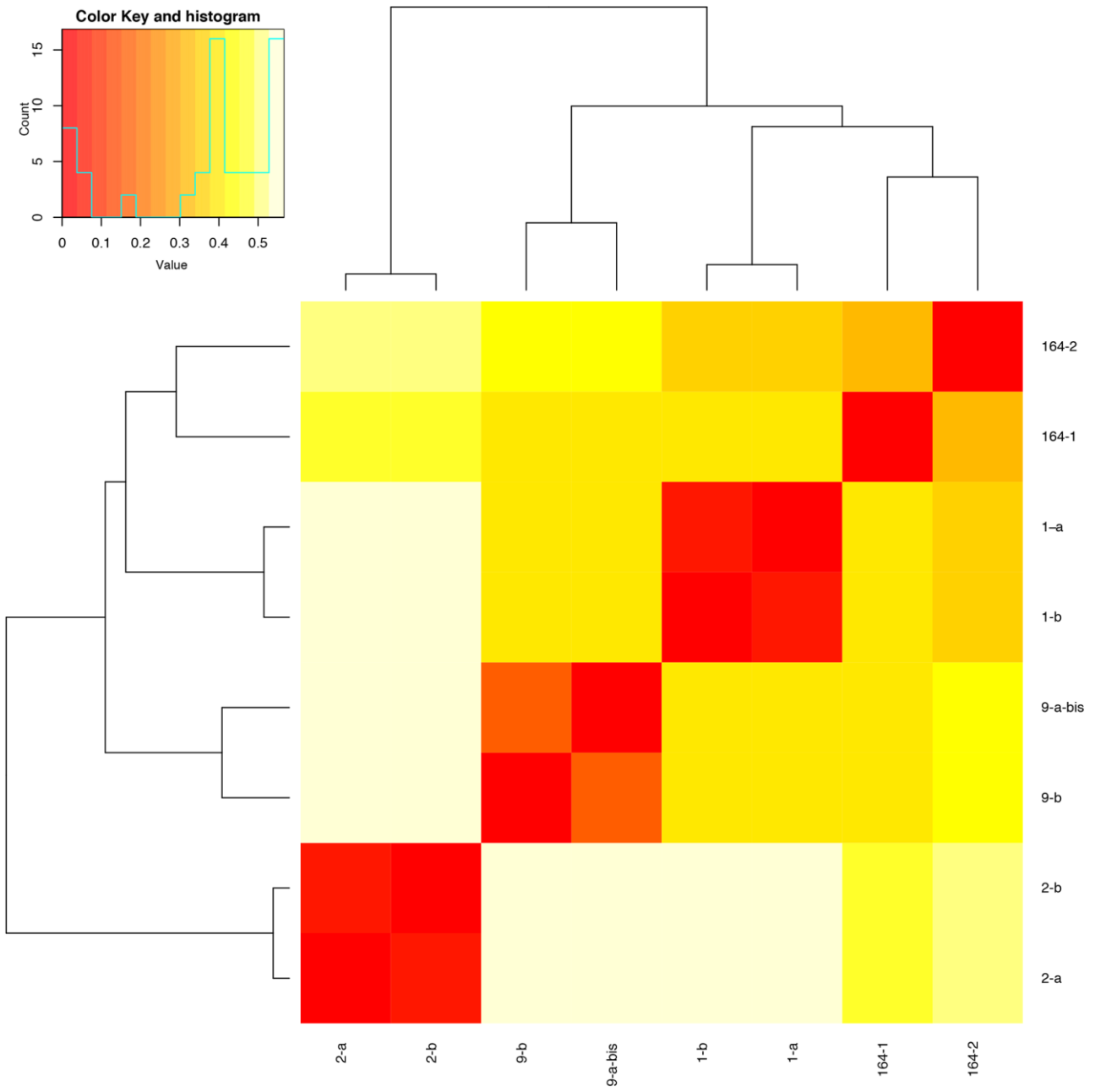


Figure 6. Heat map representing Nei's distances in all pairwise comparisons based on SNPs identified in the eight genotypes under study

Discussion

The technological revolution in NGS technology brought unprecedented opportunities to study any organism of interest both at the genomic and transcriptomic levels. RNA-seq allows the deep screening of the entire transcriptome, which includes all expressed sequence, and represents a reduced representation of the genome [29]. Apart from the analysis of global transcriptional mRNA profiling under distinct environmental or developmental conditions, which can be considered the main or original scope of this technology, RNA-seq is a powerful tool for many other applications, including the prediction of gene transcripts for those organisms without a sequenced genome, the implementation of gene predictions for those organisms already having a genome draft, the discovery of splice variants, the detection of single nucleotide variants (SNVs) and structural variations (SVs), and the profiling of small RNAs. For all these applications, transcriptome assembly is a challenging but crucial step for accurate downstream genetic analyses.

Fennel (*F. vulgare* Mill.) is a species native of the western Mediterranean areas, belonging to the family of Apiaceae (Umbelliferae). It is commonly distinguished in two sub-species: the subsp. *vulgare*, which includes the presently cultivated forms, used as vegetables or to produce aromatic fruits, and the subsp. *Piperitum* (Ucria) Cout., commonly referred to as wild fennel, widely used in central and southern Italy, to flavor different foods such as bread, cheese, fish, liqueurs, meat, meat products and salads.

Despite its agronomic and pharmaceutical interest, available molecular data on fennel are limited and only few genetic studies were performed in this species so far. The majority of gene sequences and gene products currently accessible from GenBank refers to the plastid genome which was sequenced more than 10 years ago [9]. At this regard, the present study raises from the need to develop genomic resources and molecular tools to be mined from the scientific and producer's community for future fennel molecular studies, genetic analyses and breeding initiatives.

With this aim, we used NGS technology to develop the first fennel leaf transcriptome sequencing using two different transcriptome reconstruction strategies: without the aid of a reference genome or *de novo* assembly and according to a reference genome-guided assembly. Each of the two strategies has its advantages and disadvantages, depending on the specific conditions used for the transcriptome assembly. In general, *de novo* transcriptome assembly is much more computationally expensive than genome-guided assembly, and it is used when reference genomes are not available, although they can also be utilized when genome references are available. In our strategy, we consider both a *de novo* assembly based on transcriptome sequencing (RNA-seq) conducted on four economically relevant accessions of *F. vulgare*, namely OL1, OL2, OL9 and OL164, and a genome-guided assembly performed using a low coverage reference genome draft obtained from whole-genome sequencing of the OL2 varietal genotype. Although the complete assembly of the fennel genome is beyond the scope of the current work, it is worth mentioning that the preliminary clustering of sequencing data in 300,408 scaffolds provided a very useful backbone for the genome-guided assembly of the leaf transcriptome.

As expected *de novo* transcriptome assembly produced a higher number of contigs (61,299 loci) compared to the *ex post* genome-guided assembly (51,917 loci), although the average transcript length of the latter was significantly higher than the former. Aligning the RNA-seq reads against the genome draft enabled to produce an average alignment rate of 91%, suggesting that the vast majority of the coding sequences was correctly represented in the genomic scaffolds. Moreover, the lower number of transcripts produced by the genome-guided assembly could be due to the low coverage of whole-genome sequencing. As a consequence, mapping of RNA-seq reads on short and partial genomic contigs allowed to obtain only a partial reconstruction of the fennel transcriptome. Conversely, the *de novo* transcriptome assembly, being independent of reference genome could reconstruct also transcripts not in the reference due to missing portions of the genome, structural variants or other reasons [30]. Moreover, although genome-guided assemblies still have its merits as an *ex post* transcriptome assembly, limitations imposed by spliced alignments, like errors and

artifacts, can negatively influence the results [11]. To get a higher level of reliability of the fennel transcriptome to obtain a more comprehensive transcriptome, we attempt a combined approach, combining the genome-guided and *de novo* assemblies into a clustered transcriptome. This approach led to a resulting transcriptome with a number of transcripts higher compared to both the other approaches (79,263) and with an intermediate average transcript length (1,141.94). Focusing on the combined transcriptome, we were able to annotate 47,775 (60.27%) transcripts, with a large percentage of them related to sequences from grapevine (16.88%) and sesame (10.18%). Of the 47,775 transcripts with BLASTx hits, the GO analysis determined GO ID and enzyme code (EC) assignments for 14,734 (30.8%) with full or partial annotations (see Figure 1). Of the 14,734 annotated transcripts, 1,615 have predicted functions (EC codes). Cellular metabolic processes were among the most highly represented groups in terms of GO analysis, as expected given that young leaves are undergoing rapid growth and extensive metabolic activities.

Considering the pivotal role of transcription factors in regulating many plant processes and functions, we focused on this category, identifying up to 1,011 transcripts, corresponding to the 2.1% of the total annotated transcripts identified. This result is totally in agreement with what observed in *Daucus carota*, from the same Apiaceae family [23]: in view of 57,128 annotated transcripts, the 2.9% of them (1,677) matched with as many transcription factors. The overall distribution of transcription factors in fennel within the known TF families is similar to what observed in other species including soybean [31] and chickpea [32] and, in particular, carrot [23]. The TF family most represented in our data was the bHLH, a superfamily of TFs representing important regulatory components in many transcriptional complexes, controlling processes such as regulation of flavonoid biosynthesis, epidermal cell fate determination such as stomata formation, hormone response and light signaling [33,34]. Together with bHLH the most enriched TFs in our analyses belonged to the C2H2, ERF, NAC, MYB and MYB-related families. Surprisingly, these families are the six most represented ones in *D. carota* too (see Figure 3). All the above-mentioned TFs are organized in large gene families in plants, with numbers comparable with those observed in

our study, although the low coverage of our assembly and the partial annotation of sequences make it difficult to make comparisons.

Simple sequence repeats (SSRs), or microsatellites, are largely used for genetic diversity analyses and marker-assisted breeding programs, because of their highly polymorphic and discriminant nature, co-dominant inheritance, prevalence throughout the genome, ease of use and cost-effectiveness. Being located within the coding regions, expressed SSRs (EST-SSRs) have increased amplification success in related species, are useful for assessing functional diversity and for marker-assisted selection, and can act as anchor markers for evolutionary and comparative mapping studies. As a counterpart, they possess a lower level of polymorphism compared to genomic SSR markers. The screening of the assembled transcriptome led us to identify a total number of 6,411 microsatellite regions, most of which belonging to the di- and tri-nucleotide category (for details see Figure 5). This finding is similar to what reported in celery [35], sesame [36], peach [37], kiwifruit [38], rice [39], whereas tetra-nucleotide repeats dominate in species such as bread wheat [40], grapevine [41] or sugarcane [42].

Together with the EST-SSRs, we screened out transcriptome assemblies for single nucleotide polymorphisms (SNPs), another class of markers which providing means of assessing genetic variation that, although less polymorphic than SSRs, is abundant and easily to obtain via NGS. In our study, we identified approximately 43,000 SNPs and 4,000 In/Dels among the 8 genotypes considered. The informativeness of this new set of SNPs was successfully evaluated by calculating the genetic distances in pairwise comparisons among the 8 genotypes. As expected, genetic distances between individuals from different lines were always higher (> 0.35) than those drawn from genotypes belonging to the same lines (< 0.30). Validation of the identified SNPs is not the scope of this paper, nevertheless, we believe this list presents a significant resource for future work in plant breeding and genetic diversity assessment and marks the first SNP markers discovered to date in fennel. These markers could be used in both Mendelian gene and quantitative trait loci (QTL) mapping, generating genetic linkage maps, genotyping and breeding programs. The

frequency of SNPs and In/Dels were 1 SNP every 2 Kb and 1 In/Del every 25 Kb. It should be considered that the frequency of single-nucleotide variants, such as SNPs and In/Dels, in the transcribed regions is supposed to be lower compared to non-coding regions within the genome. Finally, the transition:transversion ratio was equal to 1.65. Similar ratios were found across SNPs in other Apiaceae such as carrot, where the ratio of transition substitutions was about 1.75 to 1 [43].

The available scientific research on fennel revealed that it is an important medicinal plant used in a wide range of ethnomedical treatments, including abdominal pains, antiemetic, arthritis, cancer, diarrhea, etc. Moreover, studies carried out in the past and present indicate that fennel possesses diverse health benefits: extracts of fennel possess a range of pharmacological actions, such as antiaging, antiallergic, anticolitic, antihirsutism, anti-inflammatory, antimicrobial and antiviral [44].

Amongst the large number of chemical constituents identified in fennel, the volatile component t-anethole probably represent the most important one, both in terms of organoleptic effect, conferring the typical anise taste and in terms of medical roles. The biosynthesis of t-anethol has been recently elucidated by Koeduka *et al.* [45], which identified and characterized two genes encoding t-anol/isoegenol synthase 1 (IGS1) and t-anol/isoegenol O-methyltransferase 1 (AIMT1) in *Pimpinella anisum*. It is known that IGS1 uses coumaroyl acetate substrate to catalyze the formation of t-anol, whereas AIMT1 catalyzes the formation of t-anethole through a methylation step. The screening of the 47,775 annotated transcripts from the clustered transcriptome allowed us to identify all those genes belonging to the phenylpropanoid pathway (including PAL, PTAL, CYP73A, 4CL, CCR, CAD and CFAT, as shown in Figure 4) involved in the biosynthesis of the coumaroyl acetate compound. Moreover, we identified two transcripts encoding for genes putatively involved in the t-anethole biosynthesis. A first transcript (MSTRG.27089.1) significantly matched (E-value= $3e^{-165}$, similarity 70.5 %) with the t-anol/isochavivol synthase (IGS1, EC: 1.1.1.319) characterized in *P. anisum*, (gi|218963652) whereas a second one (MSTRG.32111.1) probably encoded (E-value=0, similarity 75.0%) for the t-anol/isochavicol O-methyltransferase (AIMT1, EC: 2.1.1.279) described by Koeduka *et al.* (gi|218963654). Considering the close

relationship between structural genes involved in the parallel pathways leading to t-anethol and methylchavicol (estragole) biosynthesis we were not able, based on the sequence identity, to discriminate whether our transcript encodes for structural genes involved in a pathway rather than the other. By the way, although it has been proved that the amount of estragole and trans-anethole produced by *F. vulgare* varied consistently during plant development [46], the production of t-anethole proved to be very high [2,47] when compared with the accumulation of estragole in this species. Taking into account this aspect, it is reasonable to think that the two transcripts aforesaid may encode for structural genes involved in the t-anethole pathway.

In conclusion, the newly assembled leaf transcriptome will represent an innovative cultural step and valuable molecular dataset exploitable for genetic and functional characterization of wild and cultivated fennel materials. We are confident that the bioinformatics characterization of the main genes and gene products involved in the t-anethole biosynthetic pathway as well as the identification of single-nucleotide variants will find soon utility for breeding new varieties in *F. vulgare*.

References

1. FAO. Food and Agriculture Organization of the United Nations: Value of Agricultural Production [Internet]. 2016 [cited 2016 May 30]. Available from: <http://faostat3.fao.org/download/Q/QV/E>
2. Díaz-Maroto MC, Pérez-Coello MS, Esteban J, Sanz J. Comparison of the volatile composition of wild fennel samples (*Foeniculum vulgare* Mill.) from Central Spain. *J Agric Food Chem*. 2006;54(18):6814–8.
3. Asano T, Aida S, Suemasu S, Mizushima T. Anethole restores delayed gastric emptying and impaired gastric accommodation in rodents. *Biochem Biophys Res Commun*. 2016;472(1):125–30.
4. Tognolini M, Ballabeni V, Bertoni S, Bruni R, Impicciatore M, Barocelli E. Protective effect of *Foeniculum vulgare* essential oil and anethole in an experimental model of thrombosis. *Pharmacol Res*. 2007;56(3):254–60.
5. Bardai S El, Lyoussi B, Wibio M, Morel N. Pharmacological evidence of hypotensive activity of *Marrubium vulgare* and *Foeniculum vulgare* in spontaneously hypertensive rat. *Clin Exp Hypertens*. 2001;23(4):329–43.
6. Senatore F, Oliviero F, Scandolera E, Tagliatela-Scafati O, Roscigno G, Zaccardelli M, et al. Chemical composition, antimicrobial and antioxidant activities of anethole-rich oil from leaves of selected varieties of fennel [*Foeniculum vulgare* Mill. ssp. *vulgare* var. *azoricum* (Mill.) Thell]. *Fitoterapia*. 2013;90:214–9.
7. Kubo I, Fujita K, Nihei K. Antimicrobial activity of anethole and related compounds from aniseed. *J Sci Food Agric*. 2008;88(2):242–7.
8. Knio KM, Usta J, Dagher S, Zournajian H, Kreydiyyeh S. Larvicidal activity of essential oils extracted from commonly used herbs in Lebanon against the seaside mosquito, *Ochlerotatus caspius*. *Bioresour Technol*. 2008;99(4):763–8.
9. Peery R, Kuehl J, Boore J, Jeffrey L, Raubeson L. Comparisons of three Apiaceae chloroplast genomes - coriander, dill and fennel. In: *Botany*, Botanical society of America. 2006.
10. Unamba CIN, Nag A, Sharma RK. Next Generation Sequencing Technologies: The Doorway to the Unexplored Genomics of Non-Model Plants. *Front Plant Sci*. 2015;6:15036.
11. Lu BX, Zeng ZB, Shi LT. Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Sci China Life Sci*. 2013;56(2):143–55.
12. Doyle J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull*. 1987;19:11–5.
13. Bolger AM, Lohse M, Usadel B. Genome analysis Trimmomatic : a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–20.
14. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2 : an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 2012;1(18):1–6.
15. Patel RK, Jain M. NGS QC Toolkit : A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS One*. 2012;7(2):e30619.
16. Kim D, Langmead B, Salzberg SL. HISAT : a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12(4):357–60.
17. Perteua M, Perteua G, Antonescu C, Chang T-C, Mendell J, Salzberg S. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2016;33(3):290–5.
18. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT : accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23):3150–2.
19. Gene Ontology Consortium. Gene Ontology Project [Internet]. 2016 [cited 2016 Dec 30]. Available

from: <http://geneontology.org/>

20. Kanehisa Laboratories. KEGG: Kyoto Encyclopedia of Genes and Genomes [Internet]. 2016 [cited 2016 Dec 30]. Available from: <http://www.genome.jp/kegg/>
21. The UniProt Consortium. UniProt : the universal protein knowledgebase. *Nucleic Acids Res.* 2017;45:158–69.
22. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinforma Oxf Engl.* 2005;21(18):3674–6.
23. Xu Z, Tan H, Wang F, Hou X. CarrotDB : a genomic and transcriptomic database for carrot. *Database.* 2014;2014:1–8.
24. Jin J, Tian F, Yang D, Meng Y, Kong L, Luo J, et al. PlantTFDB 4.0 : toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 2017;45:1040–5.
25. Thiel T, Michalek W, Varshney R, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet.* 2003;106(3):411–22.
26. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754–60.
27. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–303.
28. Nei M. *Molecular Evolutionary Genetics*. New York, NY: Columbia Press, University; 1987.
29. Huang X, Yan H-D, Zhang X-Q, Zhang J, Frazier TP, Huang D-J, et al. De novo Transcriptome Analysis and Molecular Marker Development of Two Hemarthria Species. *Front Plant Sci.* 2016;7(April).
30. Chen G, Li R, Shi L, Qi J, Hu P, Luo J, et al. Revealing the missing expressed genes beyond the human reference genome by RNA-Seq. *BMC Genomics.* 2011;12(1):590.
31. Vatanparast M, Shetty P, Chopra R, Doyle JJ, Sathyanarayana N, Egan AN. Transcriptome sequencing and marker development in winged bean (*Psophocarpus tetragonolobus*; Leguminosae). *Sci Rep.* 2016;6(1):29070.
32. Hiremath PJ, Farmer A, Cannon SB, Woodward J, Kudapa H, Tuteja R, et al. Large-scale transcriptome analysis in chickpea (*Cicer arietinum* L.), an orphan legume crop of the semi-arid tropics of Asia and Africa. *Plant Biotechnol J.* 2011;9(8):922–31.
33. Toledo-Ortiz G, Huq E, Quail PH. The Arabidopsis Basic / Helix-Loop-Helix Transcription Factor Family. *Plant Cell.* 2003;15(August):1749–70.
34. Serna L, Martin C. Trichomes: different regulatory networks lead to convergent structures. *Trends Plant Sci.* 2006;11(6):274–80.
35. Li M-Y, Wang F, Jiang Q, Ma J, Xiong A-S. Identification of SSRs and differentially expressed genes in two cultivars of celery (*Apium graveolens* L.) by deep transcriptome sequencing. *Hortic Res.* 2014;1(December 2013):10.
36. Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D, et al. Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC Genomics.* 2011;12(1):451.
37. Jung S, Abbott A, Jesudurai C, Tomkins J, Main D. Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs. *Funct Integr genomics.* 2005;5(3):136–43.
38. Fraser L, Harvey C, Crowhurst R, De Silva H. EST-derived microsatellites from *Actinidia* species and their potential for mapping. *Theor Appl Genet.* 2004;108(6):1010–6.

39. Temnykh S, Park WD, Ayres N, Cartinhour S, Hauck N, Lipovich L, et al. Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). *TAG Theor Appl Genet*. 2000;100(5):697–712.
40. Gupta P, Rustgi S, Sharma S, Singh R, Kumar N, Balyan H. Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Mol Genet genomics*. 2003;270(4):315–23.
41. Scott KD, Eggler P, Seaton G, Rossetto M, Ablett EM, Lee LS, et al. Analysis of SSRs derived from grape ESTs. *TAG Theor Appl Genet*. 2000;100(5):723–6.
42. Cordeiro G, Casu R, McIntyre C, Manners J, Henry R. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to *erianthus* and sorghum. *Plant Sci*. 2001;160(6):1115–23.
43. Rong J, Lammers Y, Strasburg JL, Schidlo NS, Ariyurek Y, de Jong TJ, et al. New insights into domestication of carrot from root transcriptome analyses. *BMC Genomics*. 2014;15(1):895.
44. Badgujar SB, Patel V V., Bandivdekar AH. *Foeniculum vulgare* Mill: A review of its botany, phytochemistry, pharmacology, contemporary application, and toxicology. *Biomed Res Int*. 2014;2014:842674.
45. Koeduka T, Baiga TJ, Noel JP, Pichersky E. Biosynthesis of t-anethole in anise: characterization of t-anol/isoeugenol synthase and an O-methyltransferase specific for a C7-C8 propenyl side chain. *Plant Physiol*. 2009;149(1):384–94.
46. Rather MA, Dar BA, Sofi SN, Bhat BA, Qurishi MA. *Foeniculum vulgare* : A comprehensive review of its traditional use , phytochemistry , pharmacology , and safety. *Arab J Chem*. 2016;9:1574–83.
47. Aprotosoia Ac, Şpac A, Hăncianu M, Miron A, Tănăsescu VF, Dorneanu V, et al. The chemical profile of essential oils obtained from fennel fruits (*Foeniculum vulgare*). *Farmacia*. 2010;58:46–53.

Supplementary materials

Figure S1. Species distribution of the top BLASTx hits for the assembled loci (E-value $\leq 1e^{-05}$)

Species	<i>N</i> transcripts	%
other species	13855	29.00
<i>Vitis vinifera</i>	8064	16.88
<i>Sesamum indicum</i>	4868	10.19
<i>Theobroma cacao</i>	2526	5.29
<i>Nicotiana tomentosiformis</i>	2281	4.77
<i>Nicotiana sylvestris</i>	2099	4.39
<i>Jatropha curcas</i>	1851	3.87
<i>Erythranthe guttata</i>	1641	3.43
<i>Solanum lycopersicum</i>	1510	3.16
<i>Solanum tuberosum</i>	1500	3.14
<i>Populus trichocarpa</i>	1363	2.85
<i>Gossypium raimondii</i>	1337	2.80
<i>Ricinus communis</i>	1323	2.77
<i>Populus euphratica</i>	1240	2.60
<i>Citrus clementina</i>	1191	2.49
<i>Prunus persica</i>	1126	2.36

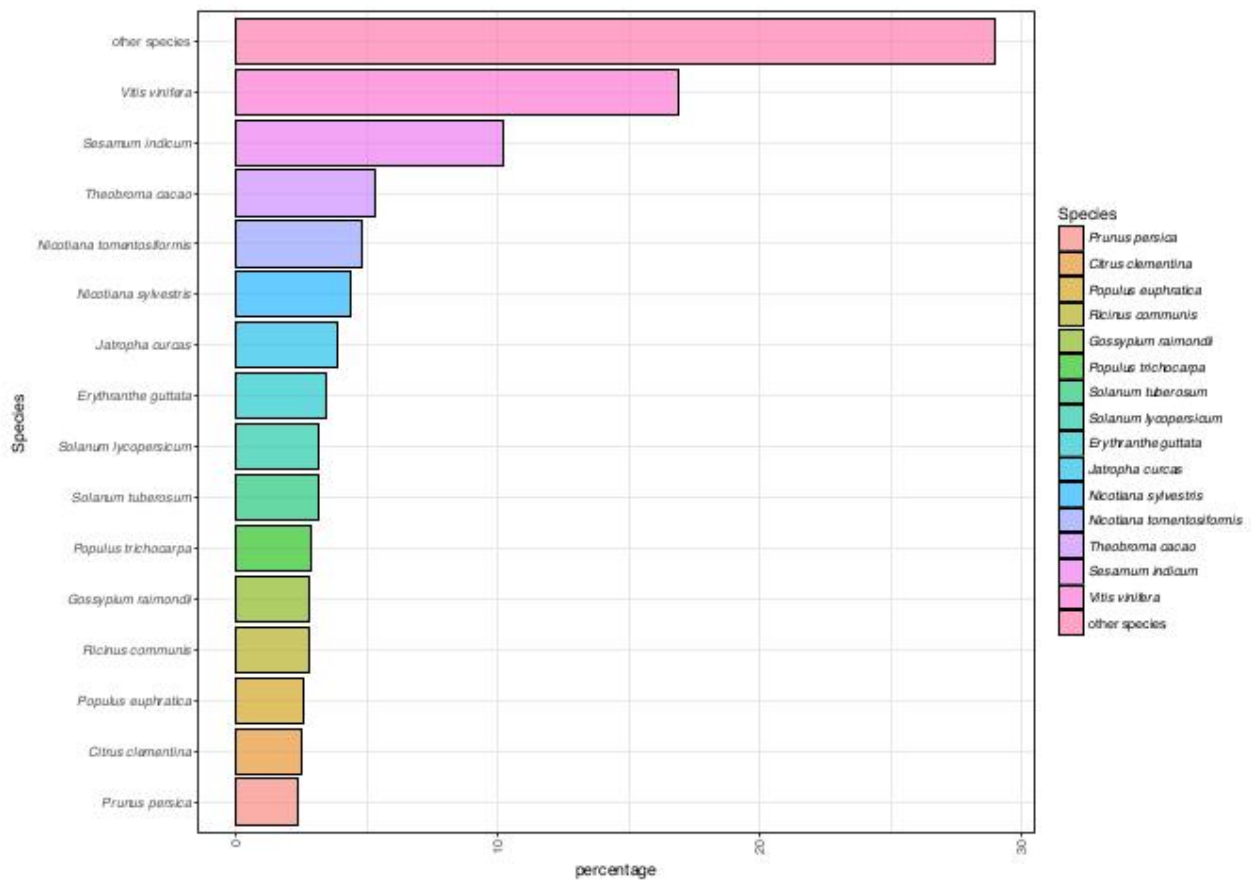


Figure S2. E-value distribution of BLASTx hits for the assembled loci

E-value range	<i>N</i> transcripts	%
0	10757	22.52
1E-100 to 1E-180	9093	19.03
1E-60 to 1E-100	7372	15.43
1E-40 to 1E-60	5488	11.49
1E-20 to 1E-40	7594	15.90
1E-5 to 1E-20	7471	15.64

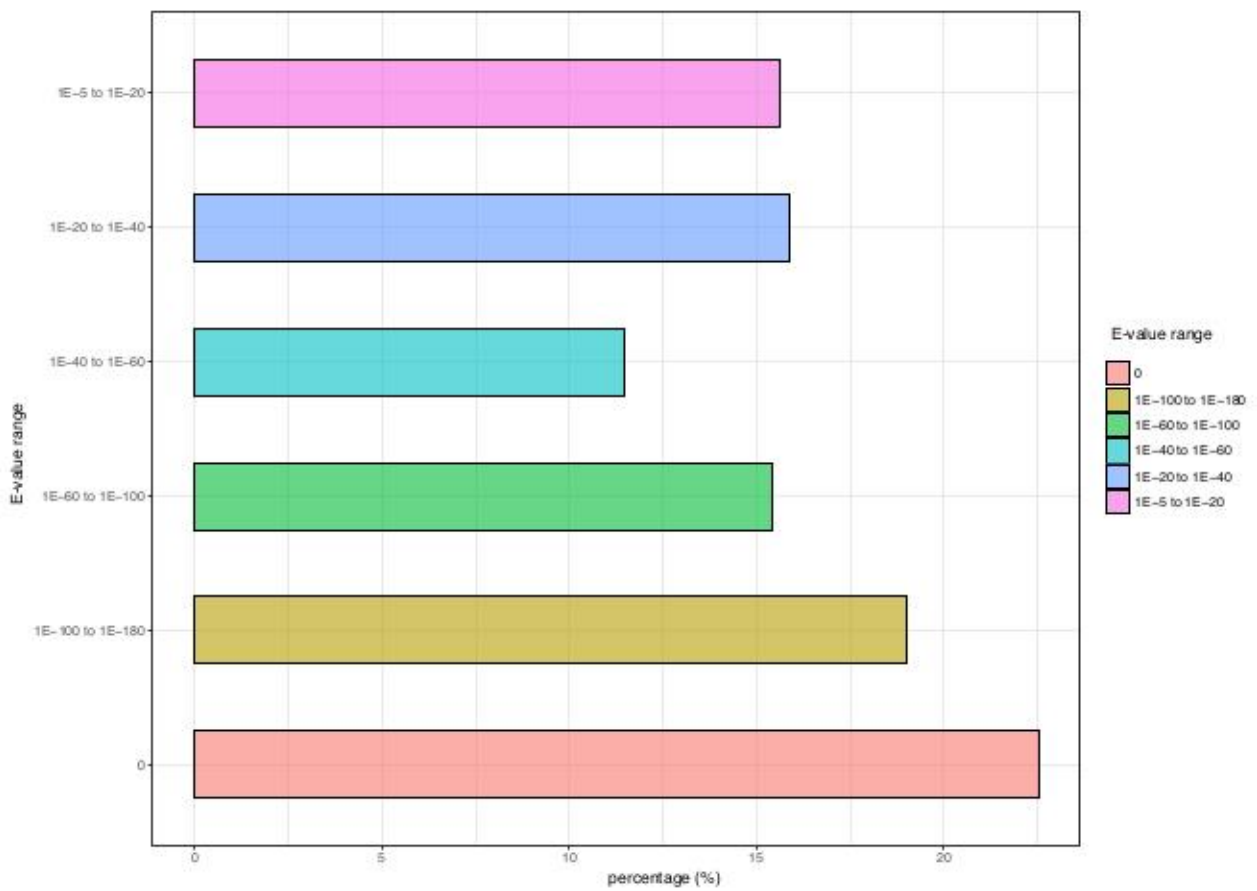
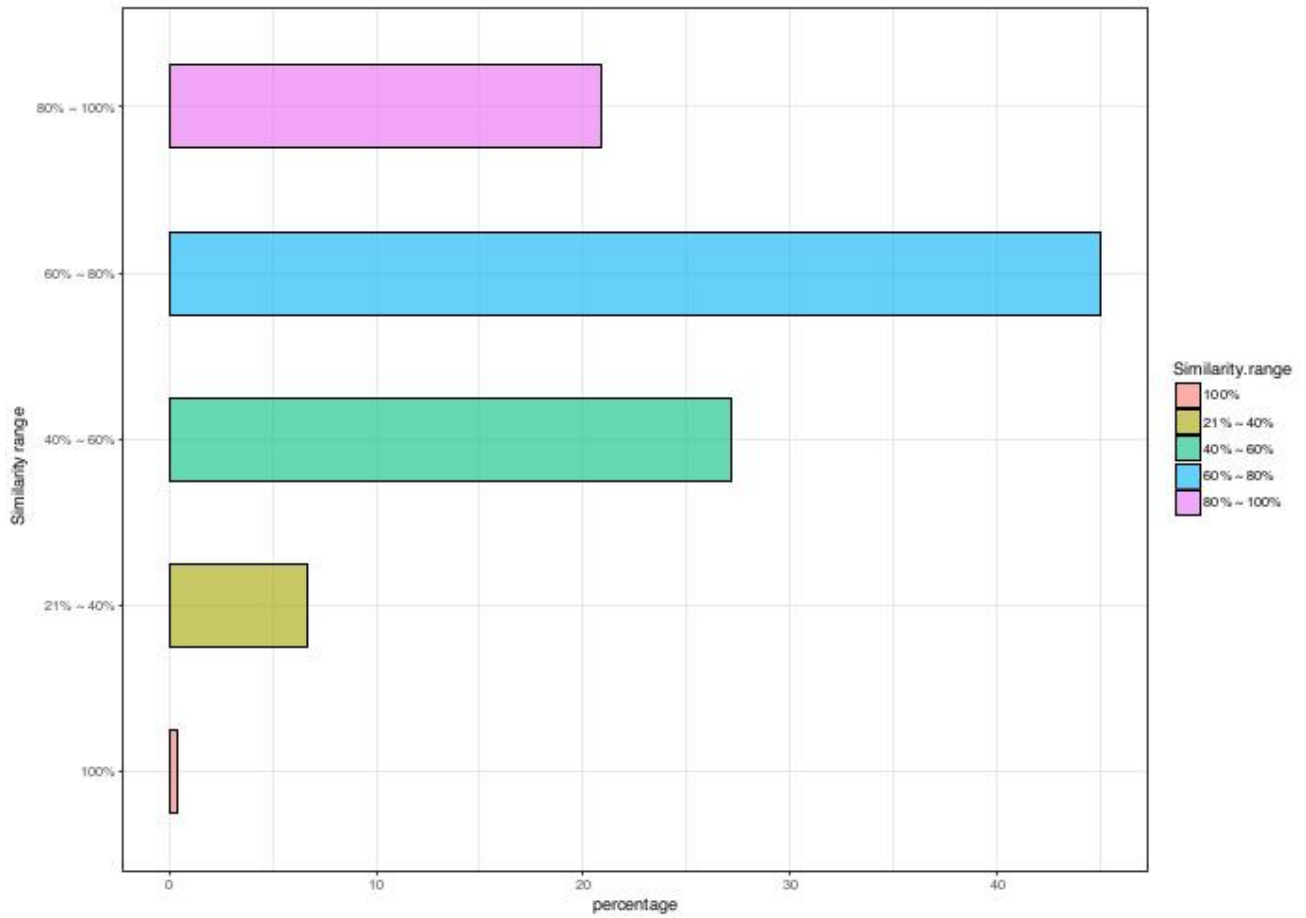


Figure S3. Similarity distribution of BLASTx hits for the assembled loci

Similarity range	<i>N</i> transcripts	%
100%	189	0.40
80% ~ 100%	9963	20.85
60% ~ 80%	21492	44.99
40% ~ 60%	12953	27.11
21% ~ 40%	3178	6.65



Chapter V

**First draft genome sequencing of fennel (*Foeniculum Vulgare* Mill.):
identification of simple sequence repeats and their application in
marker-assisted breeding**

Abstract

The development of F₁ hybrid varieties benefits from the synergistic effect of conventional and molecular marker-assisted breeding schemes. A sequencing run was carried out in *Foeniculum vulgare* ($2n=2x=22$) to develop the first genome draft and to discover microsatellites suitable for implementing multilocus SSR marker assays. A preliminary cytometric analysis allowed us to estimate the genome size ($2C=2.64-2.86$ pg), equal to about 1.34 Mbp for 1C genome, and to calculate the sequencing coverage ($53\times$). Among the 300,408 assembled scaffolds, 125,719 showed at least one significant match from a comparison against the NCBI-NR database and 71,540 loci were assigned to one or more ontological IDs. Moreover, the bioinformatic annotation enabled to detect 103,306 SSR elements. 40 microsatellites were randomly chosen among those ones with dinucleotide and trinucleotide repeat motifs and with a repeat motif length ≥ 25 times and preliminarily tested. 14 SSR markers resulted suitable for genetic diversity analyses, were efficiently organized in two PCR multiplex assays and validated using a core collection of 118 fennel individuals potentially useful for the development of inbred lines and F₁ hybrids. All SSR loci were found polymorphic, scoring an observed number of marker alleles $N_a=106$ and an average polymorphism information content $PIC=0.64$. The SSR data were used to calculate: i) the degree of homozygosity for the individual inbred lines ($0.23 < H_o < 0.92$), to eventually plan additional selfing or sibling cycles, and ii) the degree of genetic similarity for all possible pair-wise comparisons between parental inbred lines ($GS=0.19-0.54$), to identify the most divergent combinations for the constitution of experimental F₁ hybrids. The integration of genotypic and phenotypic data was useful to implement guidelines for precision hybrid breeding schemes in fennel.

Keywords: genome assembly, draft genome, sequencing, SSR markers, F₁ hybrids, fennel

Introduction

Hybrids grow bigger and stronger than their parents do. This phenomenon, known as hybrid vigor or heterosis, refers to the situation in which an F_1 progeny obtained by crossing genetically divergent inbred lines or pure lines exhibits greater biomass, resistance to biotic agents and abiotic stresses, faster development and higher fertility than the two lines used as parents [1]. In the eighteenth century, J.G. Koelreuter observed the superiority of plants resulting from interspecific crosses in various genera as *Nicotiana*, *Dianthus*, *Verbascum*, *Mirabilis* and *Datura* [2] but, officially, the concept of hybrid vigor is rooted in nineteenth century with the book “The effects of cross and self-fertilization in the vegetable kingdom”, by Charles Darwin (1876). In the first decades of the twentieth century, male-sterility was found to play an important role for the production of hybrid seed via crossing parental inbred lines appropriately selected through progeny tests. Onion (*Allium cepa* L.) represents the first species for which the hybrid effect was exploited through the aid of a male sterility system [4]. Since then, the same system was developed in different forms and in a wide range of species. In particular, cytoplasmic male sterility (CMS) proved to be far more convenient for the exploitation of the hybrid vigor than genic male sterility (GMS) because of its inheritance, mode of preservation and restoration of fertility [2]. GMS is caused by nuclear genes and it is usually governed by a single recessive gene (*ms*) with monogenic control, although a dominantly inherited pattern is also possible, whereas CMS is a maternally inherited trait that is often associated with unusual open reading frames (ORFs) found in the mitochondrial genome. The male-sterility condition, under which a plant does not generate pollen or is unable to produce functional pollen, is widespread among higher plants [5]. In the latter reproductive system, fertility can be restored only if a sterile line is crossed with a pollinator that possesses a nuclear ‘fertility restorer gene’ [6].

A fascinating biological system for the successful application and exploitation of heterosis is represented by cultivated fennel (*Foeniculum vulgare* Mill., $2n=2x=22$). In this species F_1 hybrids

are bred for their improved crop yields, the quality of the inflated leaf bases that form bulb-like structures and the resistance or tolerance to parasites and to environmental stresses. As already happened over the last decades in several crop species like for instance barley [7], corn [8], bean [9], rice [10], eggplant [11], the CMS phenotype of fennel has been widely exploited by seed companies to develop specific male-sterility lines of high breeding value. According to the most recent data available in the Food and Agriculture Organization Corporate Statistical Database [12], India is the world leader in fennel production with more than 500,000 tons per year, followed by Mexico and China. In the same year, the gross value of world fennel production was \$ 5.3 billion. Strikingly, despite the economic relevance of fennel, researchers and breeders are forced to deal with the complete lack of both biological and genomic data for this species. This aspect is particularly critical if we consider that the constitution of F₁ hybrids benefits from the synergistic effect of marker-assisted breeding (MAB) techniques and conventional breeding programs. For this reason, considering that an accurate phenotypic and genotypic evaluation of the inbred lines represents a consolidated strategy for crop improvement [13] and taking advantage of the dramatic cost reduction of the next-generation sequencing (NGS) systems, a high throughput DNA sequencing approach was carried out in *F. vulgare*. In the last few years, this rapid and cost-effective approach has been profitably applied to several non-model species including *Macadamia integrifolia* [14], *Pisum sativum* L. [15], *Carthamus tinctorius* L. [16], *Arachis hypogaea* L. [17] and *Pistacia vera* L. [18] for the development of polymorphic markers and, in particular, microsatellites. Hence, almost 40 years later their discovery, markers based on simple sequence repeat (SSR) elements still represent one of the most useful and convenient molecular tools for applied breeding purposes. The co-dominant inheritance and the high polymorphic index, make them attractive markers for genotyping aimed at the selection of best candidate lines. This information could be exploited for planning crosses and predicting both uniformity and potential heterosis of experimental F₁ hybrids based on the genetic diversity between parental lines (*i.e.*

expected heterozygosity) characterized by high levels of homozygosity for different marker alleles across several genomic loci.

The aims of this work were: i) evaluating the nuclear genome size of *Foeniculum vulgare* through flow cytometry to estimate the average coverage of the genome sequencing; ii) assembling reads produced by a sequencing run into a genome draft; iii) identifying and characterizing SSR marker loci suitable for DNA genotyping assays; iv) validating a set of SSR marker loci by performing a genome-wide characterization of a fennel core collection of lines with high breeding value.

Materials and Methods

Flow cytometry

Nuclei were isolated from 100 mg leaf tissue from three different fennel samples randomly chosen from commercial populations by gentle chopping with a razor blade in 0.4 ml of CyStain® PI Absolute P nuclei extraction buffer (Sysmex Partec GmbH, Gorlitz, Germany) supplemented with 1% w/v PVP. Nuclei suspensions were filtered by using 30 µm CellTrics® (Sysmex Partec GmbH, Gorlitz, Germany). Following the filtration step, 1.6 ml staining buffer was added to each sample and tubes were stored in the dark on ice for 1 h before measurement. The fluorescence intensity of PI-stained nuclei was determined using the flow cytometer CyFlow® Cube Ploidy Analyser (Sysmex Partec GmbH, Gorlitz, Germany) equipped with an Nd-YAG green laser ($\lambda = 532$ nm; 30 mW). PI-stained nuclei were analysed at a flow rate of 4 µl/sec using a 590 nm long pass filter. *Solanum lycopersicum* ($2C = 2.00$) and *Phaseolus vulgaris* ($2C = 1.32$) were adopted as external standards for fluorescence reference during measurements [19–21]. Each sample was analysed in triplicate. Fluorescence histograms were analysed with the FCS Express 5 Flow software (Sysmex Partec GmbH), after manual treatment to exclude noise. DNA content was inferred by comparing sample and standard G0/G1 peak positions [19].

DNA extraction, library preparation and next-generation sequencing

One sample (namely OL2) was chosen from commercial and experimental populations based on: i) commercial relevance of the cultivar; ii) robust phenotypic and genotypic characterization; iii) high degree of homozygosity (> 90%). Leaves were collected, snap-frozen in nitrogen upon harvesting and stored at -80°C until further processing. Genomic DNA was isolated using a standard CTAB protocol [22] and the integrity was assessed through an electrophoretic run (1% agarose/1× TAE gel containing 1× Sybr Safe DNA stain; Life Technologies, Carlsbad, CA, USA). Moreover, DNA concentration and purity (in terms of 260/280 nm and 260/230 nm absorbance ratios) were spectrophotometrically measured using Nanodrop2000c (Thermo Scientific, Waltham, MA, USA). About 2 μg of genomic DNA were subjected to library preparation using the Illumina TruSeq DNA PCR-free sample preparation kit (Illumina, Inc., San Diego, CA, USA) according to the instructions provided by the company. The library was sequenced on an Illumina HiSeq 2500 using paired-end, 150-bp-read chemistry (Illumina) and with an average insert size of 350 bp. Taking advantage of the genome size previously estimated through flow cytometry, raw data were used to compute the depth of sequencing through the Lander/Waterman equation [23]:

$$C = LN/G \quad (1)$$

Where C is the average coverage, L is read length, N is the number of reads and G is the haploid genome length.

Raw genomic sequences were processed with Trimmomatic software [24] to remove the adapter sequences and to trim low quality bases. In particular, Trimmomatic was run setting an average minimum quality score of 20 within a sliding window of 5 and the minimum reads length was set to 75 bp. The filtered sequences were assembled using SOAPdenovo2 [25] into contigs at distinct k-mer values (from 71 to 121). The quality of the draft assembly was assessed using NGSQCToolkit v2.3.3 [26].

The genome draft was annotated by using a subset of the NR protein database focused on the pentapetalae clade, with a BLASTX-based approach (E-value $\leq 1e^{-05}$, BLAST v.2.3.0+). Moreover, in order to extrapolate Gene Ontology annotations [27] and E.C. annotations, the GI identifiers of the BLASTX hits were mapped to the UniprotKB protein database [28].

SSR loci development and primer design

Simple sequence repeats were detected using the MicroSATellite (MISA) Identification Tool Perl script [29]. In details, the assembled sequences were screened for di-, tri-, tetra-, penta- and hexa-nucleotide repeat motifs with a minimum repeat number of 7, 6, 6, 6 and 5, respectively. The maximal number of bases interrupting two SSRs in a compound microsatellite was set at 100 and the space between imperfect SSRs at 5 bp. A total of 40 primer pairs were randomly designed in batch by using the software BatchPrimer3 v1.0 [30], by adopting the following parameters: i) dinucleotide or trinucleotide repeat motif; ii) length of the repeat motif ≥ 25 times; iii) melting temperature (T_m) always between 53°C and 56°C to facilitate their use in multiplex reactions.

Polymerase Chain Reaction (PCR) and electrophoresis

Genomic DNA samples to be used for PCR amplification of selected SSR markers was isolated by using the DNeasy 96 Plant Kit (Qiagen, Hilden, Germany).

An initial screening of a subset of three DNA samples, randomly chosen among the fennel collection, was performed to investigate the amplification efficiency of the 40 SSR primer pairs, tested in single reactions. A total of 14 SSR marker loci were then selected for the genotyping of 118 DNA samples chosen from a core collection of commercial lines and experimental materials. More in details, our investigations focused on a population composed by 45 seed parents (male sterile accessions) and 49 maintainers, overall organized in 12 sub-populations, along with 24 pollen donors (male fertile accessions) chosen from 9 different sub-populations.

The amplification reactions were carried out organizing the 14 SSR markers in two multiplex PCRs adopting the three-primer strategy reported by Schuelke [31] with some modifications. Briefly, for

each primer pair, universal sequences (namely M13 for and PAN1-3, unpublished) were used to tag the 5'-end of the forward primer and adopted in PCR reactions in combination with M13, PAN1, PAN2 and PAN3 fluorophore-labelled oligonucleotides. Fluorophores adopted in all amplification reactions were 6-FAM, VIC, NED and PET, respectively.

PCR reactions were performed in a total volume of 10 μ l containing approximately 20 ng of genomic DNA template, 1 \times Platinum[®] Multiplex PCR Master Mix (Applied Biosystems, Carlsbad, CA, USA), GC enhancer 10% (Applied Biosystems), 0.05 μ M tailed forward primer (Invitrogen Corporation, Carlsbad, CA, USA), 0.1 μ M reverse primer (Invitrogen Corporation), 0.23 μ M universal primer (Invitrogen Corporation) and sterile water to volume.

All reactions were accomplished using a 9600 Thermal Cycler (Applied Biosystems) with 96-well plates. The following thermal conditions were adopted in all reactions: 2 min at 95°C for the initial denaturing, 45 cycles at 95°C for 30 s, 55°C for 30 s and 72°C for 45 s. A final extension at 72°C for 30 min terminated the reaction, to fill-in any protruding ends of the newly synthesized strands.

Amplicons were visualized by agarose gel electrophoretic runs (agarose 2% agarose/1 \times TAE gel containing 1 \times Sybr Safe DNA stain (Life Technologies). Images of electroporetic runs were carried out with an UVITEC UV Transilluminator (Cambridge, UK) equipped with a digital camera. Fluorescent labeled PCR products were then dried at 65°C for one hour and then subjected to capillary electrophoresis, which were performed with an ABI PRISM 3130xl Genetic Analyzer (Thermo Fisher) and adopting LIZ500 (Applied Biosystems) as molecular weight standard.

Data analysis

After determining the marker allele size of each SSR locus by means of Peak Scanner 1.0 (Applied Biosystems), statistical analyses were performed using PopGene v1.32 software [32]. The observed homozygosity (H_o) for individual samples and the expected heterozygosity (H_e) within sub-populations equivalent to the unbiased Nei's genetic diversity, the Shannon Index (I) of phenotypic diversity, the number of observed alleles (N_a) and the number of effective alleles (N_e) per locus

were also calculated. In order to measure the informativeness of SSR markers, the polymorphism information content (PIC) was calculated for each locus by using the following formula with the Excel Microsatellite Tool Kit [33]:

$$PIC=1-\sum_{i=1}^n p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^n 2p_i^2 p_j^2 \quad (2)$$

Genetic similarity estimates for all possible pair-wise comparisons among 118 fennel samples were calculated with NTSYS v2.2 software [34], applying the simple matching coefficient.

Results

Genome size estimates

Genome size determination was performed by using flow cytometry (FCM) of propidium iodide-stained nuclei. Fennel samples were co-stained with nuclei extracted either from *S. lycopersicum* [20] or *P. vulgaris* [19], which were adopted as internal reference standards, and the relative fluorescence was used to calculate the genome size of fennel. The two reference standards used, cultivated tomato and common bean, were selected as they provided similar, but not fully overlapping fluorescence histograms (Figure 1), two conditions that are believed to favour precise genome size estimation by limiting instrumental nonlinearity FCM errors [35,36].

All FCM histograms showed two predominant peaks, corresponding to the G0/G1 nuclei of the assessed samples (Figure 1). Our estimates of the 2C genome size, which were calculated by comparing the relative G0/G1 nuclei peak of propidium iodide-fluorescence corresponding to *F. vulgare* with those recorded for *S. lycopersicum* and *P. vulgaris*, ranged from 2.64 pg to 2.86 pg. Noteworthy, estimates of the genome size by using two reference standards selected among different taxonomic families provided highly overlapping values.

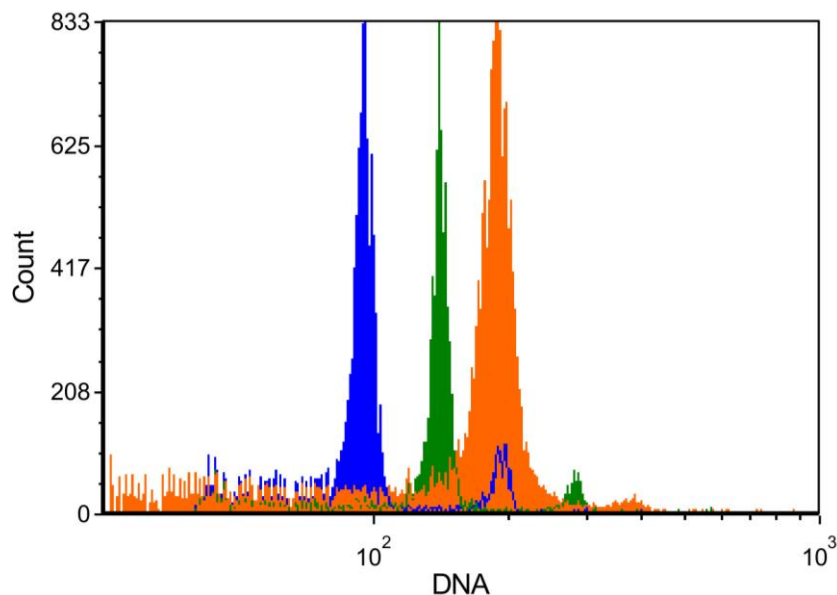


Figure 1. Flow cytometry analysis for genome size estimates. Histograms represent the total DNA fluorescence emission of propidium iodide-stained leaf nuclei purified from *F. vulgare* (orange), *S. lycopersicum* (green) and *P. vulgaris* (blue)

Assembly of the genome draft and development of multiplex SSR primer sets

A total of 486,073,396 paired end reads, corresponding to 72.91 Gbp, were generated by means of an Illumina HiSeq 2500 platform. After the trimming step, considering the size estimate of the haploid genome ($C=1.32-1.43$ Gbp) and applying the Lander/Waterman equation, the average coverage was found to be approximately $53\times$. The assembly resulted highly fragmented and split into as many as 7,978,334 contigs ($N50=319$). Considering the assembled scaffolds, we obtained a total of 300,408 sequences whose length ranged from 370 to 145,787 bp for a total of 1.01 Gbp ($N50=9,443$, Table 1). Therefore, the Illumina sequencing allowed the assembly of about 75% of the whole estimated genome of fennel. All sequences of the assembled scaffolds were deposited as Whole Genome Shotgun (WGS) project at GenBank under the accession PHNY000000000.

From BLASTX analysis ($E\text{-value} \leq 1e^{-05}$) performed against the “pentapetalae clade” subset of the NR protein database, 125,719 scaffolds (41.85% of the total number) showed at least one significant match. Among them, 15,743 exhibited similarity scores higher than 80%. Scaffolds showing a BLASTX match were imported in Uniprot for GO mapping and EC annotation. A total of 71,540 loci were assigned to at least one ontological ID and 3,465 EC numbers were ascribed to as many scaffolds.

Table 1. Main descriptive statistics related to the assembled data of fennel genome draft

Main Statistics	Genome
Total sequences	300,408
Total bases	1,011,093,015
Min sequence length	370
Max sequence length	145,787
Average sequence length	3,365.73
Median sequence length	1,241.00
N25 length	20,734
N50 length	9,443
N75 length	2,842
N90 length	1,126
N95 length	777
As %	31.81
Ts %	31.72
Gs %	16.06
Cs %	16.11
(A+T)s %	63.53
(G+C)s %	32.17
Ns %	4.30

Predetermined MISA scripts allowed the identification of 103,306 SSRs distributed over 50,631 scaffolds. On the whole, 85.90% were perfect SSRs, 0.34% imperfect SSRs and 13.76% compound SSRs. The most frequent repeats were dinucleotides (69.80%) and trinucleotides (27.30%), as shown in Figure 2, while the most abundant dinucleotide and trinucleotide repeat motif types were AT/AT (45.71%) and AAT/ATT (23.74%), respectively (Table 2).

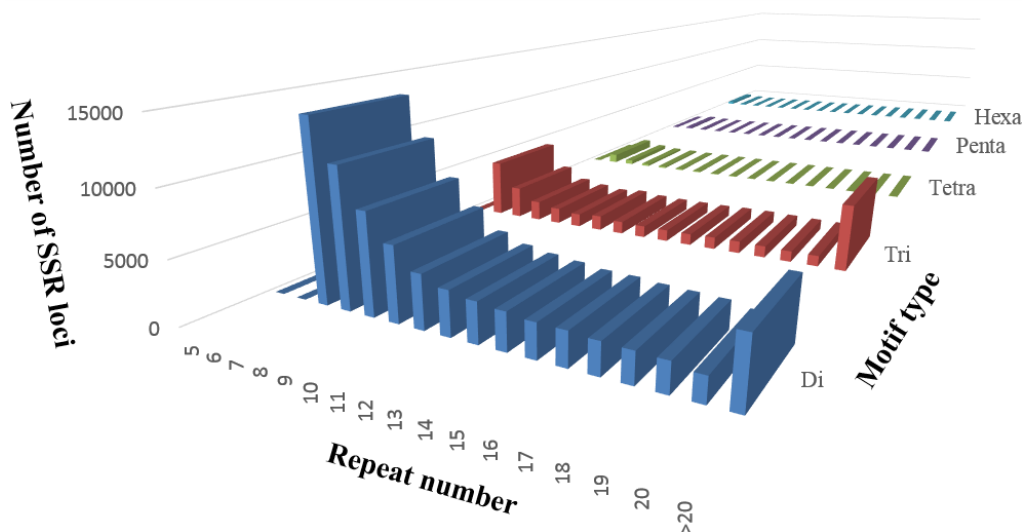


Figure 2. Number of SSR loci detected, categorized by motif types and repeat numbers (from 5 to >20) in the *de novo* assembled genomic sequences of *Foeniculum vulgare*

Table 2. Frequency of classified repeat types (considering sequence complementary) in fennel

Repeats	<10	10	11	12	13	14	15	16	17	18	19	20	>20	Total	%
AT/AT	18282	3836	2892	2524	2346	2210	2076	2141	2042	1965	1925	1611	3376	47226	45.71
AG/CT	7137	1025	679	519	458	411	387	322	288	287	266	243	1569	13591	13.16
AC/GT	7479	1009	640	451	347	304	212	165	122	113	84	61	183	11170	10.81
CG/CG	75		2										0	77	0.07
AAT/ATT	8300	1010	1195	997	968	951	947	1004	1014	966	906	879	5391	24528	23.74
AAG/CTT	1071	40	42	30	14	15	16	15	15	11	5	12	94	1380	1.34
AAC/GTT	514	50	34	39	13	23	17	14	10	5	11	6	44	780	0.76
ACAT/ATGT	915	83	89	56	48	31	28	26	13	7	1	6	20	1323	1.28
Others														3231	3.03

Overall, 10,086 SSR regions distributed over 9,821 scaffolds resulted not exploitable for primer design because the motif was detected within the first/last 50 bases of the scaffold. The remaining SSR regions were used to design specific primer pairs for 40 SSR loci, randomly selected among those encompassing a dinucleotide or a trinucleotide repeat motif and showing a repeat motif length higher than 25 nucleotides. Then these SSR loci were tested using three genomic DNA samples arbitrarily selected from the complete set of available accessions (Table 3).

The BLASTX analysis revealed that 22 SSR loci were located in as many scaffolds scoring significant matches with proteins included in the GeneBank NR database. More in details, considering a 5 kb long genomic window from the target SSR, 15 loci of these were located upstream or downstream of as many predicted genes, whereas the remaining 7 loci were positioned within predicted genes. Of the 40 SSR-specific primer pairs, two failed the amplification of discrete PCR products for single loci, while as many as 24 loci revealed non-specific amplification with multiple PCR products. The remaining 14 loci, being amplified efficiently and generating unambiguous and polymorphic profiles, were organized in two multiplex PCR assays (Figure 3).

Table 3. Information on the microsatellite markers isolated in *Foeniculum vulgare*, including locus name, primer sequence, motif, amplicon size, marker localization with respect to the closest genic region, RefSeq ID and predicted function of the closest genic region. The 14 SSR loci written in bold were organised in multiplex and validated in this study on a commercial collection of 118 fennel individuals

Name primer	Sequence	Size	Motif	Marker localization	RefSeq ID	Predicted function
FV_179608	<u>TCTACTTTTAATTCCCCATC</u> <u>GATACACGTATAGTTAAGAAGCA</u>	338	TCT	Downstream	XP_004499217.1	DNA-directed RNA polymerase II subunit rpb2-like
FV_489	<u>ACAACCTCAACAGCAAACAGT</u> <u>GTTGTTGTGATTGGAATATGAG</u>	222	TAT	-	-	-
FV_51455	<u>TATCACGGGCAAGTCTATTTA</u> <u>TATTTTAGTTGCCCAAAGCTA</u>	164	AAT	-	-	-
FV_79693	<u>CGTCTTAATTGAAACGGATAA</u> <u>TATCACGGGCAAGTCTATTTA</u>	225	TAT	-	-	-
FV_51379	<u>TAAAAATATCTCGGGCAAGTA</u> <u>TCCGTGATATGAGAAAGGTAA</u>	185	AAT	Upstream	XP_012851407.1	Uncharacterized protein
FV_237	<u>AAATGCCCTAAAAATACCTT</u> <u>GTGCAATACTAAGCCTTTTGA</u>	286	AAT	-	-	-
FV_6	<u>TATGTTCTCAGATTCGGGTTA</u> <u>GTTCATCAAACCTGTGTCATTGT</u>	186	TC	-	-	-
FV_25339	<u>CTGCTCTGAATCCACAAATA</u> <u>CCCTAAACAATCACAAAAATG</u>	274	GA	Upstream	YP_009179710.1	Photosystem II protein D1
FV_144120	<u>CTCTTTTCCAAAAATATCACG</u> <u>GATGAAAAAGGGTAATTGGTT</u>	162	AAT	-	-	-
FV_288620	<u>AGGTTCAACCAAATTATACCC</u> <u>CAGGTGTTCTTCTGATTATG</u>	221	ATT	-	-	-
FV_462	<u>ATGGCTGAGAATTAGGGTTAC</u> <u>CGATCTACGCCTTAGAGGTAT</u>	186	AG	Upstream	XP_010668292.1	Uncharacterized protein
FV_255981	<u>ACGTGACTAAACAAGAGATGC</u> <u>GTATTTGTTATTCGATGAATGTG</u>	240	TAA	-	-	-
FV_253	<u>TTGTAGAGATACAGGGTCGAA</u> <u>GAGGGGAGTCAGTTAAACAAC</u>	192	TC	Upstream	NP_001276240.1	GDP-D-mannose pyrophosphorylase
FV_144370	<u>GAGCAATCAAAACATCTCATT</u> <u>GTAACATGTTTTGGAAAGAAAGA</u>	280	AAT	-	-	-
FV_217218	<u>ACAAACGTACCTCTGTACGAA</u> <u>TCAGAAGGTGAGTTATGTTGC</u>	292	AG	Downstream	XP_011080831.1	Cation/H(+) antiporter 24
FV_51427	<u>ACGGGGTGTTTAAAAATGTAAT</u> <u>CTATTTTCCAAAAATCTCACG</u>	186	TAT	-	-	-
FV_255590	<u>AAATTGATCCGAAACTAAACC</u> <u>ATAGTGGACCGACAATGTAAC</u>	302	TAA	Downstream	XP_009781234.1	Replication factor A protein 1-like
FV_79739	<u>GGGCAATTAATAATCTCAGG</u> <u>AGTTGTCCGTGATATGAGAAA</u>	186	AAT	-	-	-
FV_288738	<u>TGTTGAGGTTGATTATTGTTAGA</u> <u>CACCTCGAAACCTTAGTGTTGA</u>	296	TAT	Upstream	XP_011014816.1	Uncharacterized protein
FV_79504	<u>TCGGGCGTTACATTATTATT</u> <u>TTTCACTAAAACACCCAAAA</u>	194	TAT	-	-	-
FV_179831	<u>GGCAATCAAAACATCTTCTTT</u>	283	AAT	Upstream	XP_012087681.1	Uncharacterized

	TAAATAGACTTGGCCGTGATA						protein
FV_92	<u>ATCTCCGTTTTTAAGGACAAC</u> <u>TCCGTGATATGAGAAAGGTAA</u>	202	AAT	-	-	-	-
FV_179566	<u>CGGTAAATCTAATCACAACG</u> <u>CACGGCTACAAAGCTAGT</u>	315	TA	-	-	-	-
FV_65	<u>GCCTATGTATTTGCAAGAATG</u> <u>TGCAACATTCAATTGTGTAGA</u>	199	CT	Downstream	XP_010319957.1		Uncharacterized protein
FV_25097	<u>AAAACGTTCCCTAAGTTTTGAG</u> <u>GGTTCAGAAGAAATTTCCAAG</u>	252	ATA	Upstream	XP_010693752.1		Uncharacterized protein
FV_110142	<u>GCGAATTCAACGGATAGATA</u> <u>TATTTACAGCACTCTTTTGCTC</u>	199	AAT	Upstream	XP_008455983.1		KH domain-containing protein
FV_217360	<u>TGCTCATGATATGAGAATGGT</u> <u>ACGGGCAACTAAAATATCTCT</u>	254	TTA	-	-	-	-
FV_51426	<u>CCTGAAACACTTCAAATCAAA</u> <u>AAAACAGGGACTCCATAGAAC</u>	201	TAA	Downstream	XP_008347142.1		Uncharacterized protein
FV_217225	<u>AAAGAATGGAGAGAAGAATGG</u> <u>ACTAGAAATAGGGGTACGTG</u>	300	AG	Upstream	XP_010671205.1		Uncharacterized protein
FV_289009	<u>CCAAAAATATCACGGACAAGT</u> <u>TTGTGTATGTTGGGAAAAGT</u>	203	TAA	-	-	-	-
FV_179837	<u>ATTCACCATGACATCACCTC</u> <u>ACAGTGTGGGTTTGTATGTGT</u>	305	TC	Downstream	XP_007024187.1		Cysteine-rich receptor-like protein kinase
FV_11537	<u>TTCATGTATCAACTACGCACAC</u> <u>CTCTGGGATTGGATTCAAGGAG</u>	141	AG	Included	XP_008365831.1		dTDP-4-dehydrorhamnose reductase-like
FV_28045	<u>TTAGAATAAGGGTAGGGCAACGG</u> <u>ATGCAATTTAACACTGTGGTGTGG</u>	239	AC	-	-	-	-
FV_15981	<u>CTAGCGTTTCCATCTCGTCTC</u> <u>AACCCGTAACTTTAACCACCAC</u>	213	TC	Included	XP_004491003.1		Sorting nexin 2A-like
FV_16201	<u>CCTCCATCTATTTGTGGACGA</u> <u>GAGAATTGAGGAAGAAAGCGAG</u>	93	TC	Included	XP_017252893.1		Mannosyl-oligosaccharide 1,2-alpha-mannosidase
FV_18902	<u>GTTTGAACCTCGAATGACCACCT</u> <u>GGGTCTATCATCACTCTCGC</u>	390	TC	Included	XP_009626305.1		zinc finger CCCH domain-containing protein
FV_2349	<u>AATCAATGGATGTTTGTATGAG</u> <u>GAAGAGACTTTGACTGGCATA</u>	148	AG	Included	XP_002283810.1		COP9 signalosome complex subunit 2 isoform X2
FV_3665	<u>ACACCTAGCATCACAAGGCA</u> <u>ATCAGAATCTGGGATTAGTTTGGG</u>	190	TG	Included	XP_002516152.1		GTP-binding protein, alpha subunit,
FV_9919	<u>AGTAAAGGCATAATCTGTTGGTGG</u> <u>TCATATTATCAACCTCAGGCACAG</u>	209	GT	Included	XP_010693544.1		Uncharacterized protein
FV_2	<u>CAAAGAATGGAAAACATGCTG</u> <u>CTTTCCATTGTCAATTTGC</u>	126	(CAA)	-	-	-	-

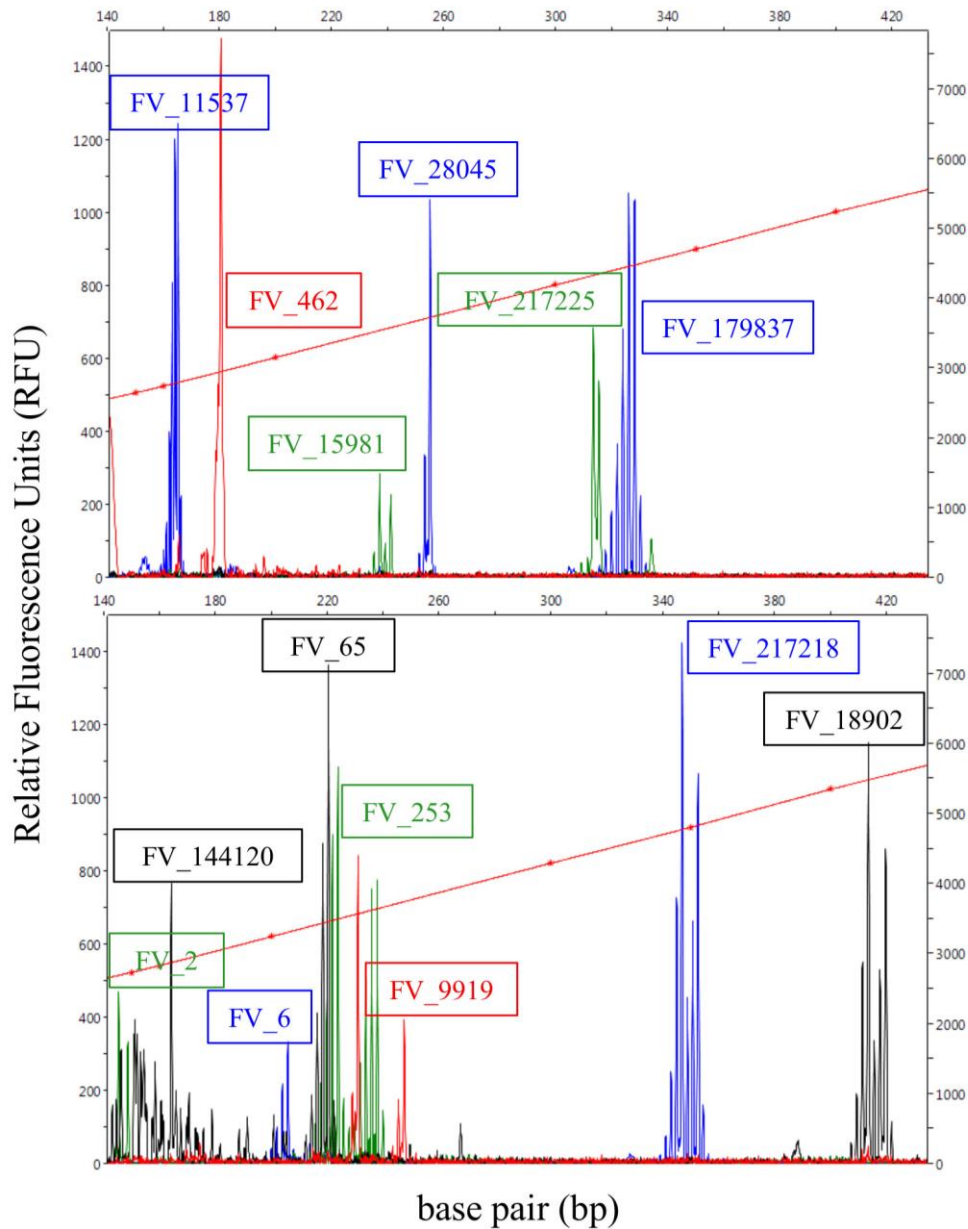


Figure 3. Electropherograms related to multiplex SSR sets analysis: Multiplex 1 (6 loci, upper panel) and Multiplex 2 (8 loci, lower panel)

Marker-assisted breeding in fennel

The new set of 14 SSR marker loci was validated through the genetic characterization of 118 fennel breeding stocks, including male-sterile inbred lines, with the related maintainer accessions, and male-fertile inbreds or pollinators. The PIC values of these SSR marker loci varied from 0.02 (FV_462) to 0.86 (FV_6), with an average value equal to 0.64 (Table 4).

Table 4. Global genetic diversity statistics sorted by locus related to the samples (n=118) of fennel (*Foeniculum vulgare*) and calculated using 14 SSR loci. Included parameters are the observed homozygosity (Ho), the unbiased Nei's genetic diversity equivalent to the expected heterozygosity (He), the number of observed (Na) and effective (Ne) alleles. The estimates of Shannon's information index of phenotypic diversity (I) and the polymorphism information content (PIC) coefficient were also calculated for each locus

Locus	Ho	He	Na	Ne	I	PIC
FV_462	0.98	0.02	2.00	1.02	0.05	0.02
FV_11537	0.70	0.72	8.00	3.53	1.52	0.68
FV_15981	0.52	0.67	7.00	3.01	1.45	0.64
FV_28045	0.77	0.34	3.00	1.51	0.56	0.29
FV_179837	0.53	0.80	8.00	5.04	1.80	0.78
FV_217225	0.72	0.75	6.00	3.96	1.49	0.71
FV_144120	0.74	0.86	10.00	6.79	2.03	0.84
FV_6	0.36	0.87	12.00	7.73	2.21	0.86
FV_217218	0.34	0.83	10.00	5.70	1.91	0.80
FV_65	0.62	0.70	10.00	3.29	1.65	0.68
FV_18902	0.39	0.81	8.00	5.24	1.82	0.79
FV_9919	0.60	0.56	6.00	2.25	1.07	0.50
FV_2	0.53	0.65	4.00	2.82	1.18	0.59
FV_253	0.38	0.84	12.00	6.13	2.03	0.82
Mean	0.58	0.67	7.57	4.14	1.49	0.64
St. Dev	0.19	0.24	3.13	2.01	0.60	0.24

As a whole, 106 marker alleles were obtained, ranging from 2 to 12 per locus (on average, 7.6). The loci FV_144120 (Na=10), FV_6 (Na=12), FV_217218 (Na=10), FV_65 (Na=10) and FV_253 (Na=12) were characterized by the highest number of marker alleles. The observed homozygosity computed across all SSR marker loci and individual DNA samples ranged from 0.34 (locus FV_217218) to 0.98 (locus FV_462), while the expected heterozygosity ranged from 0.87 (locus FV_6) to 0.02 (locus FV_462). The main genetic diversity statistics are summarized on Table 4.

It is worth mentioning that the mean and highest estimates of observed homozygosity (Ho) were calculated also for each of the sub-populations, including male-sterile lines with related maintainers and male-fertile accessions or pollen donors (Table 5). As an example, the highest values for seed parents and pollen donors were equal to 0.73 and 0.92, respectively, supporting a relatively high genetic uniformity and stability of the corresponding lines/accessions (Table 5).

The level of genetic differentiation existing among lines/accessions was investigated primarily by calculating genetic similarity estimates for all possible pair-wise comparisons among the 118 DNA individual samples. Genetic similarity (GS) values within fennel materials ranged from 0.06 to 1.00, with an average value of 0.39. The average similarity values calculated in all possible pairwise

comparisons between male-sterile lines, maintainers and male-fertile accessions are reported in Table 5. When calculated within each sub-population, these similarity values varied from 0.62 to 1.00. In the pairwise comparisons between male-sterile lines (including related maintainers) and male-fertile accessions, the average similarity values ranged from a minimum of 0.19 (between Pop1 and MfN) and a maximum of 0.54 (between Pop9 and MfF). This information is potentially useful for selecting the most genetically divergent and antagonist combinations for the development of highly heterozygous F₁ hybrids. Similarly, this information may be crucial also to avoid cross combinations between genetically similar parental materials in order to prevent inbreeding depression.

Table 5. Average genetic similarity estimates (SM) for all possible pairwise comparisons among 12 sub-populations (Pop1-Pop12) - each constituted by a variable number of male-sterile individuals and related maintainers - and male-fertile (MfA-MfN) accessions are reported along with the average observed homozygosity value within each sub-population (Mean Ho). Samples with the highest Ho value (Max Ho) within each population are reported too

				Sample name	MfA-B08	MfD-B06	MfE-B08	MfF-C10	MfG-H06	MfI-D08	MfL-F03	MfM-F11	MfN-H09
				Max Ho	0.64	0.64	1.00	0.77	0.71	0.79	0.92	1.00	0.92
				Mean Ho	0.64	0.51	0.62	0.77	0.70	0.54	0.80	0.85	0.92
				SM	1.00	0.75	0.85	1.00	0.82	0.81	0.93	0.80	0.98
Sample name	Max Ho	Mean Ho	SM		MfA	MfD	MfE	MfF	MfG	MfI	MfL	MfM	MfN
Ms1-A07	0.77	0.65	0.80	Pop1	0.37	0.27	0.27	0.31	0.32	0.27	0.22	0.34	<u>0.19</u>
Ms2-C06	0.65	0.50	0.73	Pop2	0.35	0.41	0.43	0.40	0.38	0.40	0.24	0.43	0.27
Ms3-E07	0.77	0.73	0.73	Pop3	0.32	0.36	0.46	0.44	0.41	0.35	0.31	0.39	0.28
Ms4-G07	0.64	0.55	0.70	Pop4	0.28	0.51	0.41	0.40	0.41	0.47	0.25	0.36	0.32
Ms5-D05	0.71	0.54	0.63	Pop5	0.33	0.42	0.39	0.52	0.41	0.39	0.40	0.35	0.41
Ms6-G07	0.47	0.46	0.64	Pop6	0.29	0.42	0.38	0.36	0.34	0.46	0.24	0.33	0.30
Ms7-A02	0.44	0.23	0.83	Pop7	0.30	0.35	0.48	0.46	0.32	0.39	0.35	0.38	0.34
Ms8-B03	0.43	0.37	0.62	Pop8	0.38	0.40	0.41	0.47	0.35	0.40	0.34	0.37	0.39
Ms9-B11	0.65	0.51	0.71	Pop9	0.28	0.47	0.36	<u>0.54</u>	0.27	0.39	0.37	0.36	0.28
Ms10-C08	0.72	0.60	0.81	Pop10	0.37	0.39	0.45	0.43	0.46	0.35	0.31	0.34	0.45
Ms11-E05	0.43	0.43	0.62	Pop11	0.38	0.42	0.38	0.46	0.34	0.33	0.34	0.38	0.30
Ms12-G05	0.72	0.60	0.75	Pop12	0.24	0.48	0.37	0.47	0.32	0.38	0.27	0.31	0.24

Discussion

Although fennel is a crop species that plays an important role in the nearly worldwide food culture for its agronomic, nutritional and pharmaceutical properties, researchers and breeders are forced to deal with the complete lack of biological and genomic data for this species. The main goal of the present research was to develop an informative panel of simple sequence repeat (SSR) loci to be used for breeding applications in *F. vulgare*. Because of their multi-allelic and co-dominant nature, and owing to their reproducibility among laboratories, this marker system is excellent for genetic diversity analysis (*e.g.* [37]) Mendelian gene and QTL mapping (*e.g.* [38]) and the construction of linkage maps (*e.g.* [39]).

Taking advantage of the dramatic cost reduction of the next-generation sequencing (NGS) systems, an Illumina HiSeq 2500 sequencing was performed in fennel. Before starting any sequencing experiment, understanding the depth of the sequencing achievable through a single run represents a crucial step for the robustness of the sequencing and, in particular, the reliability of polymorphism detection [40]. The depth of the sequencing or theoretical coverage is defined as the average number of times that each nucleotide is expected to be sequenced assuming a certain number of randomly distributed reads of a given length [41]. Since the genome size is required to calculate this index and the only reference available in the scientific literature for *F. vulgare*, predicted through spectrophotometer analysis, is dated 28 years ago [42], a flow cytometry analysis was performed. Unlike earlier analytical methods, flow cytometry analyses microscopic particles in suspension, which are ideally forced to flow in single file within a fluid stream through the focus of intense light. Among all tested protocols for genome size estimation by using flow cytometry, PI staining has the advantage that it is relatively fast, works with a wide variety of materials and provides information on a very large number of nuclei. According to Bennett *et al.* [43] a potential limitation of PI staining is that inhibitors and variability in chromatin condensation can bias the result, by forcing sub-optimal or variable access of PI into the major groove of DNA. This limitation can be

minimized by suitable choice of run conditions, the most important of which is the use of an internal standard that is co-prepared with the sample. Furthermore, according to Dolezel and Bartos (2005), to avoid instrumental nonlinearity FCM errors, an ideal DNA reference standard should have a genome size that is similar but not identical to the unknown sample. According to the genome size of our selected references, the mean value of the *F. vulgare* 2C DNA content range from 2.64 to 2.86 pg. Noteworthy, estimates of genome size by using two reference standards selected among different taxonomic families provided highly overlapping values. According to Bennett *et al.* [43], variability among replicated measurements, which is the order of about 7% of the estimated 2C DNA content (0.22/2.86), could be due to the presence inhibitors preventing PI binding to DNA or variability in chromatin condensation across samples.

The depth of fennel sequencing (53×) resulted to be entirely comparable to the ones achieved in other recent studies focused on SSR development [17,18] and in some cases even higher [16].

Based on our current assembly, the genome draft of fennel is made up of approximately 300,000 scaffolds, covering 1.01 Gbp. According to the total DNA content of somatic cells of this species (2C=2.64-2.86 pg), the estimated genome size is approximately 1.34 Mbp (assuming 1 pg = 0.978 Gbp; [44] and hence our genome draft covers more than 75% of the whole genome sequence. Complete assembly of the fennel genome is beyond the scope of the current work, since the low sequence coverage, the usage of a single DNA library and the presence of repetitive DNA elements prevents successful assembly. Nevertheless, a preliminary assembly of sequencing data provided useful information for microsatellite identification. The comparison between the total number of scaffolds and the number of SSR-containing scaffolds (50,631) highlighted that a significant part of the genome is characterized by repeated elements. Furthermore, it is likely that filters adopted in this study to select SSR loci suitable for MAB practices led to underestimate the abundance of repetitive elements within the genome. However, based on our data, two clarifications are needed. First, the total length of the SSR-containing motifs detected in the genome draft (3.21 Mbp) barely represents the 0.32% of the assembled sequences proving that, despite their universal distribution,

they represent, numerically, only a small portion of the whole genome. Second, 9,821 SSR-containing scaffolds (19.40% of the total) were not usable to design primers since the microsatellite motif were retrieved within the first or the last 50 bases of the scaffold. This finding seems to confirm that the genome assembler fails in repetitive regions, probably due to the shortness of the reads. A supplementary long reads sequencing approach could be useful to bypass this gap [45]. Considering the motif abundance, not surprisingly, the dinucleotide motifs were the most represented type of repeats (69.8%) followed by trinucleotide motif (27.3%), as already reported in a comprehensive review surveying 100 studies [46].

The 40 SSR marker loci suitable for DNA genotyping were randomly selected throughout the first genome draft on the basis of two firm criteria: i) di- and tri-nucleotide motifs were preferred because of their abundance and ease scoring due to the stuttering effect, which can prove to be particularly helpful at the practical level in order to distinguish artefacts from true alleles; ii) loci with more than 25 repetitions were taken into account since higher mutation rates are correlated with high number of mutations [47,48].

The BLASTX strategy performed against the non-redundant (NR) pentapetalae protein database allowed to predict the potential association of the microsatellites with expressed regions. The fact that, overall, the 55% of the 103,306 SSR loci and, specifically, 22 out the 40 validated loci were closely located upstream or downstream annotated genes or even within annotated genes, could be extremely valuable for future marker-assisted selection (MAS) purposes.

Following the design of primers, each primer pair was individually tested in singleplex PCRs on three DNA samples randomly chosen among the available fennel sample collection. The fact that only 14 of the 40 primer pairs (35%) amplified properly and specifically, generating polymorphic amplicons, unambiguous and easily detectable profiles, could be due to the high complexity-nature of these loci that makes more likely the possibility of errors during the assembly. This is consistent with the findings reported for other species: increasing the number of SSR loci investigated, the percentage of microsatellite exploitable remains usually low. Among the most recent studies, only

204 (21%) of the 950 primer couples selected in *Pistacia vera* [18], produced polymorphic and easily scorable SSR loci, almost half (49%) of the 1,644 SSR investigated in *Pisum sativum* was unusable [15] and only the 18% of 1,637 putative SSR markers analysed in *Arachis hypogaea* L, were suitable for genetic analysis [17]. The synthesis and validation tests of hundreds of primers heavily affect the total cost of the analysis, nullifying all the benefits deriving from the dramatic cost reduction of the next-generation sequencing of the last years. Comparing this study with previously published results, the number of primers tested and, thus, the cost of the analysis performed were approximately from 10 to 20 times lower. Despite that, 14 SSR loci were found to be efficiently polymorphic and highly discriminant, enabling to approach the genetic diversity resilient in our plant collection with the purpose of MAB in fennel. A future goal will be the integration of the current markers with additional markers in order to make SSR-based genotyping investigations more robust and informative. Realistically, based on our findings, testing 80-100 couples of primers specific to as many SSR loci should be enough to identify a very informative set of 20-30 SSR loci, possibly 2 or 3 per linkage group.

The 14 couples of primers, organized in two SSR multiplex (Figure 3) to further reduce the total cost of the analysis, were validated on 118 DNA samples from a fennel core collection of high breeding value with the aim of selecting the optimal parental lines for the development of F1 hybrids.

The development of F1 hybrid varieties in fennel guarantees: i) maximized crop yields; ii) high plant uniformity; iii) possibility to combine useful traits from two highly genetically divergent parental lines into their offspring. This is usually achieved by taking advantage of the cytoplasmic male sterility and through a targeted and programmed phenotypic and genotypic characterization of parental individuals. Typically, high levels of homozygosity for two parental lines characterized by high genetic diversity, one pollen donor and one seed parent (the latter, a male sterile accession nearly isogenic to its maintainer), represents the main prerequisites for the development of F1 hybrids with high predicted heterozygosity (potentially associated to heterotic vigour). In details, 12

sub-populations, each one constituted by male-sterile lines and their isogenic maintainers, were analysed and compared with 9 sub-populations of male-fertile individuals.

All the 14 SSR loci used to genotype the 118 DNA samples proved to be discriminant and polymorphic, and so informative for MAB purposes (Table 4). PIC has become the most widely applied formula for genetic studies to measure the polymorphic feature of molecular markers and it is defined as the probability that the marker genotype of a given offspring will allow deduction, in the absence of crossing over, of which of the two marker alleles of the affected parents it received [49]. According to Botstein *et al.* [50], 12 out of 14 SSR loci resulted to be highly informative ($PIC > 0.5$), one was reasonably informative ($0.5 > PIC > 0.25$), and only one was slightly informative ($PIC < 0.25$). It is extremely likely that the PIC and number of alleles per locus (N_a) could be considerably higher switching the target of the analysis from modern varieties to local varieties of different geographical origin. The overall observed heterozygosity (calculated as $1-H_o$) was on average equal to 42%, consistent with the allogamous reproductive system of this species. Nevertheless, the fact that for the fennel materials analyzed in this study the observed homozygosity (on average $H_o=58\pm 19\%$) over the 14 assayed loci was always higher than the expected one (on average $H_e=33\pm 24\%$) confirms an excess of homozygosity due to the adoption of inbreeding strategies, like repeated selfing and sibling cycles. This surplus is in agreement with the type of experimental populations (*i.e.* inbred lines), which in turn is functional to the goal of maximizing the heterozygosity rate in the resulting F1 hybrids in order to eventually exploit heterosis and to guarantee progenies with the desired genetic stability and, consequently, phenotypic uniformity.

By crosschecking the homozygosity analysis with the genetic similarity estimates for all possible pair-wise comparisons among 118 accessions, it was possible to develop a detailed guideline to organize the breeding activities, including selfing, sibling and crossing schemes (Table 5).

Low levels of genetic similarity scored within some sub-populations (*e.g.* Pop8 and Pop11) proved a scarce isogenicity between the male-sterile lines (Ms8 or Ms11) and their related maintainers (Mt8 and Mt11, respectively) and were likely a symptom of pollen contamination occurring during

the breeding program. In this case, it would be necessary to intervene on two fronts: i) improving the isolation system of these populations (to avoid further contamination); ii) performing two more cycles of backcrossing between the male-sterile accession and the related maintainer to increase the isogenicity. Overall, it was possible to predict some promising crossbreeds. For example, a cross between Pop1 and MfN was highly suggested since both scored a high level of similarity within each line (80% and 98%, respectively) and, on average, a very low level of similarity among lines (19%). In particular, we recommended the use of Ms1-A07 (male sterile from Pop1) and MfN-H09 (pollen donator from MfN population) since characterized by high level of homozygosity (77% and 92%, respectively). We are confident that by crosschecking all the genotypic data generated from this study and integrating them with phenotypic data collected in the field, it would be possible to find matches for the development of F1 hybrids.

In conclusion, our study provided an insight into the development of thousands of useful microsatellite markers through a cost-effective genome sequencing. In particular, the first genome draft (53× coverage) available for *F. vulgare* enabled the detection of 103,306 SSR regions and the implementation of an SSR assay based on 14 loci suitable for successful genotyping of fennel accessions. This information could be exploited for planning crosses and predicting plant vigour traits (i.e., heterosis) of experimental F1 hybrids on the basis of the genetic distance and allelic divergence between parental inbred lines. Knowing the parental genotypes would allow not only to protect newly registered varieties but also to assess the genetic purity and identity of the seed stocks of commercial F1 hybrids, and to certificate the origin of their food derivatives.

References

1. Barcaccia G, Falcinelli M, Lorenzetti S. Sull'eterosi nelle piante: dall'ipotesi genetica di Jones all'era genomica. Perugia U of, editor. Perugia; 2006. 112 p.
2. Budar F, Pelletier G. Male sterility in plants : occurrence , determinism , significance and use. *Life Sci.* 2001;324:543–50.
3. Darwin C. The effects of cross and self fertilisation in the vegetable kingdom. Murray J, editor. London; 1876. 471 p.
4. Jones H, Clarke A. Inheritance of male sterility in the onion and the production of hybrid seed. *Proc Am Soc Hortic Sci.* 1943;43:189–94.
5. Hanson MR, Bentolila S. Interactions of Mitochondrial and Nuclear Genes That Affect Male Gametophyte Development. *Plant Cell.* 2004;16:154–70.
6. Black M, Bewley J, Halmer P. *The Encyclopedia of Seeds: Science, Technology and Uses.* Bewley J, Halmer P, editors. CABI; 2006. 828 p.
7. Ahokas H. Cytoplasmic Male Sterility in Barley Cytoplasmic Male Sterility in Barley. *Acta Agric Scand.* 1979;29(3):219–24.
8. Dewey R, Timothy D, Levings C. A mitochondrial protein associated with cytoplasmic male sterility in the T cytoplasm of maize. *PNAS.* 1987;84(15):5374–8.
9. Mackenzie S, Pring D, Bassett M, Chase C. Mitochondrial DNA rearrangement associated with fertility restoration and cytoplasmic reversion to fertility in cytoplasmic male sterile *Phaseolus vulgaris* L . *Proc Natl Acad Sci.* 1988;85:2714–7.
10. Wang Z, Zou Y, Li X, Zhang Q, Chen L, Wu H, et al. Cytoplasmic Male Sterility of Rice with Boro II Cytoplasm Is Caused by a Cytotoxic Peptide and Is Restored by Two Related PPR Motif Genes via Distinct Modes of mRNA Silencing. *Plant Cell.* 2006;18(March):676–87.
11. Saito T, Matsunaga H, Saito A, Hamato N, Koga T, Suzuki T, et al. A Novel Source of Cytoplasmic Male Sterility and a Fertility Restoration Gene in Eggplant (*Solanum melongena* L .) Lines. *J Japan Soc Hort Sci.* 2009;78(4):425–30.
12. FAO. Food and Agriculture Organization of the United Nations: Value of Agricultural Production [Internet]. 2016 [cited 2016 May 30]. Available from: <http://faostat3.fao.org/download/Q/QV/E>
13. Jiang G-L. Molecular marker-assisted breeding: a plant breeder's review. In: Al-Khayri J, Jain S, Johnson D, editors. *Advances in Plant Breeding Strategies: Breeding, Biotechnology and Molecular Tools.* 2015. p. 431.472.
14. Nock CJ, Elphinstone MS, Ablett G, Kawamata A, Hancock W, Hardner CM, et al. Whole Genome Shotgun Sequences for Microsatellite Discovery and Application in Cultivated and Wild Macadamia (Proteaceae). *Appl Plant Sci.* 2014;2(4):1300089.
15. Yang T, Fang L, Zhang X, Hu J, Bao S, Hao J, et al. High-Throughput Development of SSR Markers from Pea (*Pisum sativum* L .) Based on Next Generation Sequencing of a Purified Chinese Commercial Variety. *PLoS One.* 2015;10(10):e0139775.
16. Ambreen H, Kumar S, Variath MT, Joshi G. Development of Genomic Microsatellite Markers in *Carthamus tinctorius* L . (Safflower) Using Next Generation Sequencing and Assessment of Their Cross-Species Transferability and Utility for Diversity Analysis. *PLoS One.* 2015;10(8):e0135443.
17. Zhou X, Dong Y, Zhao J, Huang L, Ren X, Chen Y, et al. Genomic survey sequencing for development and validation of single-locus SSR markers in peanut (*Arachis hypogaea* L .). *BMC Genomics.* 2016;17:420.
18. Motalebipour EZ, Kafkas S, Khodaeiaminjan M, Çoban N, Gözel H. Genome survey of pistachio (*Pistacia vera* L .) by next generation sequencing : Development of novel SSR markers and genetic diversity in *Pistacia* species. *BMC Genomics.* 2016;1–14.

19. Arumuganathan K, Earle ED. Nuclear DNA Content of Some Important Plant Species. *Plant Mol Biol Report*. 1991;9(3):208–18.
20. Obermayer R, Leitch I, Hanson L, Bennett M. Nuclear DNA C-values in 30 Species Double the Familial Representation in Pteridophytes. *Ann Bot*. 2002;90:209–17.
21. Praca-Fontes M, Carvalho C, Clarindo W, Cruz C. Revisiting the DNA C-values of the genome size-standards used in plant flow cytometry to choose the “ best primary standards .” *Plant Cell Rep*. 2011;30:1183–91.
22. Doyle J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull*. 1987;19:11–5.
23. Lander ES, Waterman S. Genomic Mapping by Fingerprinting Random Clones : A Mathematical Analysis. *Genomics*. 1988;2:231–9.
24. Bolger AM, Lohse M, Usadel B. Genome analysis Trimmomatic : a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–20.
25. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2 : an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 2012;1(18):1–6.
26. Patel RK, Jain M. NGS QC Toolkit : A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS One*. 2012;7(2):e30619.
27. Gene Ontology Consortium. Gene Ontology Project [Internet]. 2016 [cited 2016 Dec 30]. Available from: <http://geneontology.org/>
28. The UniProt Consortium. UniProt : the universal protein knowledgebase. *Nucleic Acids Res*. 2017;45:158–69.
29. Thiel T, Michalek W, Varshney R, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet*. 2003;106(3):411–22.
30. You FM, Huo N, Gu YQ, Luo M, Ma Y, Hane D, et al. BatchPrimer3 : A high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics*. 2008;9:253.
31. Schuelke M. An economic method for the fluorescent labeling of PCR fragments A poor man ’ s approach to genotyping for research and high-throughput diagnostics . *Nat Biotechnol*. 2000;18:233–4.
32. Yeh F, Rong-cai Y, Boyle T, Freeware MW. POPGENE, the user friendly shareware for population genetic analysis. Molecular Biology and Biotechnology Centre, University of Alberta. Alberta, Canada; 1997.
33. Park S. The Excel microsatellite toolkit. Dublin: Animal Genomics Laborator, University College; 2001.
34. Rohlf FJ. NTSYS-pc: Numerical Taxonomy and Multivariate Analysis System, version 2.0. Port Jefferson, New York: Applied Biostatistics Inc; 2008. p. 37.
35. Dolezel J, Bartos J. Plant DNA Flow Cytometry and Estimation of Nuclear Genome Size. *Ann Bot*. 2005;95:99–110.
36. Greilhuber J, Temsch EM, Loureiro JCM. Nuclear DNA Content Measurement. In: *Flow Cytometry with Plant Cells: Analysis of Genes, Chromosomes and Genomes*. 2007. p. 67–101.
37. Bohra A, Jha R, Pandey G, Patil PG, Saxena RK. New Hypervariable SSR Markers for Diversity Analysis , Hybrid Purity Testing and Trait Mapping in Pigeonpea [*Cajanus cajan* (L .) Millspaugh]. *Front Genet*. 2017;8(March):377.
38. Ohyama A, Shirasawa K, Matsunaga H, Negoro S. Bayesian QTL mapping using genome - wide SSR markers and segregating population derived from a cross of two commercial - F 1 hybrids of tomato. *Theor Appl Genet*. 2017;130(8):1601–16.

39. Dhaka N, Mukhopadhyay A, Paritosh K. Identification of genic SSRs and construction of a SSR-based linkage map in *Brassica juncea*. *Euphytica*. 2017;213:15.
40. Sims D, Sudbery I, Illott NE, Heger A, Ponting CP. Sequencing depth and coverage : key considerations in genomic analyses. *Nat Rev*. 2014;15(2):121–32.
41. International Human genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001;409:860–921.
42. Das A, Mallick R. Variation in karyotype and nuclear DNA content in different varieties of *Foeniculum vulgare* Mill. *Cytologia (Tokyo)*. 1989;54:129–34.
43. Bennett M, Price H, Johnston J. Anthocyanin Inhibits Propidium Iodide DNA Fluorescence in *Euphorbia pulcherrima* : Implications for Genome Size Variation and Flow Cytometry. *Ann Bot*. 2008;101:777–90.
44. Bennett MD, Leitch IJ. Nuclear DNA amounts in angiosperms: Progress, problems and prospects. *Ann Bot*. 2005;95(1):45–90.
45. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *PNAS*. 2011;108(4):1513–8.
46. Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, et al. Current trends in microsatellite genotyping. *Mol Ecol Resour*. 2011;11(4):591–611.
47. Petit R, Deguilloux M, Chat J, Grivet D, Garnier-Gere P, Vendramin G. Standardizing for microsatellite length in comparisons of genetic diversity. *Mol Ecol*. 2005;14:885–90.
48. Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res*. 2008;18:30–8.
49. Nagy S, Poczai P, Cernák I, Gorji AM, Hegedus G, Taller J. PICcalc: An online program to calculate polymorphic information content for molecular genetic studies. *Biochem Genet*. 2012;50(9–10):670–2.
50. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*. 1980;32(3):314–31.

Chapter VI

Construction of the first SNP-based linkage map using genotyping-by-sequencing and mapping of the *ms1* male-sterility gene in leaf chicory

Abstract

We report the construction of the first SNP-based linkage map in leaf chicory (*Cichorium intybus* subsp. *intybus* var. *foliosum*, $2n=2x=18$) and the map location of a recessive nuclear gene responsible for male-sterility (*msl*). Male-sterility is widely exploited in leaf chicory breeding programs, as it is one of the most effective methods to develop F₁ hybrid varieties. A BC₁ population of leaf chicory, segregating for the male sterility trait (1 *MsImSl* : 1 *msImSl*), was generated by crossing a male-sterile mutant of the cultivated type with a male-fertile plant of the wild form of chicory and then used for the identification of molecular markers tightly linked to the *msl* locus. From an initial SSR-based approach using 198 BC₁ progeny plants, it was possible to identify the linkage group carrying the *msl* locus with two SSR markers that were found genetically linked to the target gene at 5.8 cM and 12.1 cM apart. Then a Genotyping-by-Sequencing (GBS) was used to produce a high-density SNP-based linkage map containing 727 genomic loci. The SNP loci were organized into nine linkage groups and spanned a total length of 1,413 cM. A total of 128 coding regions were finely located throughout the genetic map, anchoring the leaf chicory genome draft to the mapped reads by means of BLASTN alignments. Most importantly, 13 Thirteen SNPs were tightly linked to the *msl* locus in the set of 44 progeny analyzed. The position of these SNPs relative to *msl* was validated using allele-specific PCR assays in an additional 64 progeny, enabling to verify that four of them co-segregated with the *msl* gene. According to the functional annotation of the genomic reads/contigs-carrying these mapped SNPs, it was possible to elect a transposase of the MuDR family as candidate gene for the *ms* trait. Moreover, the locus responsible for sporophytic self-incompatibility (*SSI*) in leaf chicory was successfully located on linkage group 5 by comparative mapping, with three putative cysteine-rich receptor-like protein kinases that were mapped in the surrounding genomic region. On the whole, this molecular information could find utility and be practically exploited for genotyping parental inbred lines and for developing commercial F₁ hybrid varieties in leaf chicory through marker-assisted breeding schemes.

Keywords: *Cichorium intybus*, genetic linkage map, male sterility, *ms1* locus, single nucleotide polymorphism (SNP) markers, genotyping-by-sequencing (GBS)

Introduction

Linkage maps based on molecular markers play a key role in the study of the genetics and genomics of crop plants. Among the possible applications, the development of high-density linkage maps has simplified the discovery of Mendelian genes [1–3] and quantitative trait loci (QTL) [4–6]. The first genetic linkage map of chicory (*Cichorium intybus* subsp. *intybus* L.), a leafy vegetable crop belonging to the family Asteraceae and widely cultivated in many European countries, consisted of 431 SSR and 41 EST markers, and covered 878 cM [7].

Chicory is a diploid plant species ($2n=18$) that is naturally allogamous, due to an efficient sporophytic self-incompatibility system [8,9]. In addition, outcrossing is promoted by a number of traits, including: i) a floral morpho-phenology (*i.e.* proterandry, with the anthers maturing before the pistils) unfavourable to selfing in the absence of pollen donors [10,11]; and ii) a competitive advantage of allo-pollen grains and tubes (*i.e.* pollen genetically diverse from that produced by the seed parents) [12,13]. Two main botanical varieties can be recognized within *C. intybus* subsp. *intybus* to which all the cultivated types of chicory belong. The first is var. *foliosum*, which traditionally includes Witloof chicory, Pain de sucre, Catalogne and Radicchio and all the cultivar groups whose commercial products are the leaves (*i.e.* leaf chicory). The second is var. *sativum* and comprises all the types whose commercial product, either destined to industrial transformation or direct human consumption, is the root (*i.e.* root chicory). In root chicory, Gonthier *et al.* [14] identified molecular markers associated with the Nuclear Male-Sterility 1 (NMS1) locus and the Sporophytic Self-Incompatibility (SSI) locus. These two loci were both mapped to narrow genomic regions belonging, respectively, to linkage groups 5 and 2 of the genetic map developed by Cadalen *et al.* [7]. Similarly, in leaf chicory, Barcaccia and Tiozzo [15,16] mapped molecular markers linked to the male-sterility gene (*ms1*) within linkage group 4, according to the map by Cadalen *et al.* [7]. Recently, a chicory genetic linkage map spanning 1,208 cM was developed by Muys *et al.* [17] using an F₂ population composed of 247 plants. This map comprised 237 markers (*i.e.* 170 AFLP,

28 SSR, 27 EST-SNP and 12 EST-SSR markers) and covered about 84% of the chicory genome. The markers were then used to find potential orthologs based on sequence homology in mapped lettuce EST clones from the Compositae Genome Project Database [17]. A total of 27 putative orthologous pairs were retained, pinpointing seven potential blocks of synteny that covered 11% of the chicory genome and 13% of the lettuce genome, opening new avenues for the comparative analysis of these two species.

Mapping of the self-incompatibility and male-sterility mechanisms in chicory is important, not only to understand the genetic basis of the main reproductive barriers that act in flowering plants, but also because of the potential applications of these loci for breeding F₁ hybrid varieties. In fact, although in the past chicory varieties were mainly synthetics produced by intercrossing a number of phenotypically superior plants, selected on the basis of morpho-phenological and commercial traits, recently private breeders and seed firms have developed methods for the development of F₁ hybrids. In the last century, male sterile mutants have allowed the exploitation of heterosis (*i.e.* hybrid vigour) through the development of F₁ hybrid varieties in many agricultural and horticultural crops. In general, male-sterility is defined as the failure of plants to develop anthers or to form functional pollen grains and it is more prevalent than female-sterility. In nature, male sterile plants have reproduction potentials because they can still set seeds, as female-fertility is unaffected by most of the mutations responsible for male-sterility. This behavior is known to occur spontaneously via mutations in nuclear and/or cytoplasmic genes involved in the development of anthers and pollen grains. Barcaccia and Tiozzo [15,16] have recently identified and characterized a spontaneous male sterile mutation in cultivated populations of leaf chicory, namely Radicchio (*Cichorium intybus* subsp. *intybus* var. *foliosum* L.). Cytological analyses revealed that microsporogenesis proceeds regularly up to the development of tetrads when the microspores arrest their developmental program showing a collapse of the exine. At the beginning of microgametogenesis, non-viable shrunken microspores were clearly visible within anthers. Moreover, genetic segregation data derived from

replicated F₂ and BC₁ populations clearly supported a nuclear origin, monogenic control and recessive nature of the male-sterility trait in the leaf chicory mutants [15,16].

In this work, taking advantage of the method of genotyping based on 27 mapped microsatellite marker loci scattered throughout the linkage groups of leaf chicory [18] and the first genome sequence draft of leaf chicory with the functional annotation of more than 18,000 unigenes [19], we successfully constructed a high-density linkage map and finely mapped the *ms1* locus in leaf chicory. After a preliminary genetic mapping of the *ms1* locus using SSR and EST markers, a Genotyping-By-Sequencing (GBS) methodology was used to narrow down the chromosomal window around the *ms1* gene, first of all developing well-saturated linkage groups for this species and then selecting molecular markers and candidate genes for male-sterility exploitable for marker-assisted breeding and gene cloning programs.

Materials and Methods

Plant materials and genomic DNA extraction

Several male sterile mutants sharing the same genotype at the *ms1* locus were discovered within local varieties of radicchio “Red of Chioggia” stemmed from recurrent phenotypic selection programs [15,16]. A backcross (BC₁) population segregating 1:1 for the male-sterility trait and comprising 198 individual plants was generated as follows. A male-sterile mutant plant (genotype *msms*), belonging to a cultivated population of radicchio was crossed with a male-fertile plant (genotype *MsMs*) selected from a spontaneous accession of wild chicory, in order to maximize genetic diversity and polymorphism levels. An F₁ male-fertile plant heterozygous at the male-sterility locus (genotype *Msms*) was selfed and an F₂ male-sterile progeny (genotype *msms*) was then backcrossed as seed parent to a sister F₁ male-fertile plant (genotype *Msms*) used as pollen donor (Figure 1). At flowering, all individuals of the BC₁ population were phenotyped using three flowers per plant by visual observations of the anthers (*in vivo* screening for the presence/absence of pollen) and cytological investigations (*in vitro* staining of pollen using acetocarmine and DAPI

solutions). Microscopy characterized a mutant phenotype by shapeless, small and shrunken microspores as compared to the wild-type ones (Figure 1).

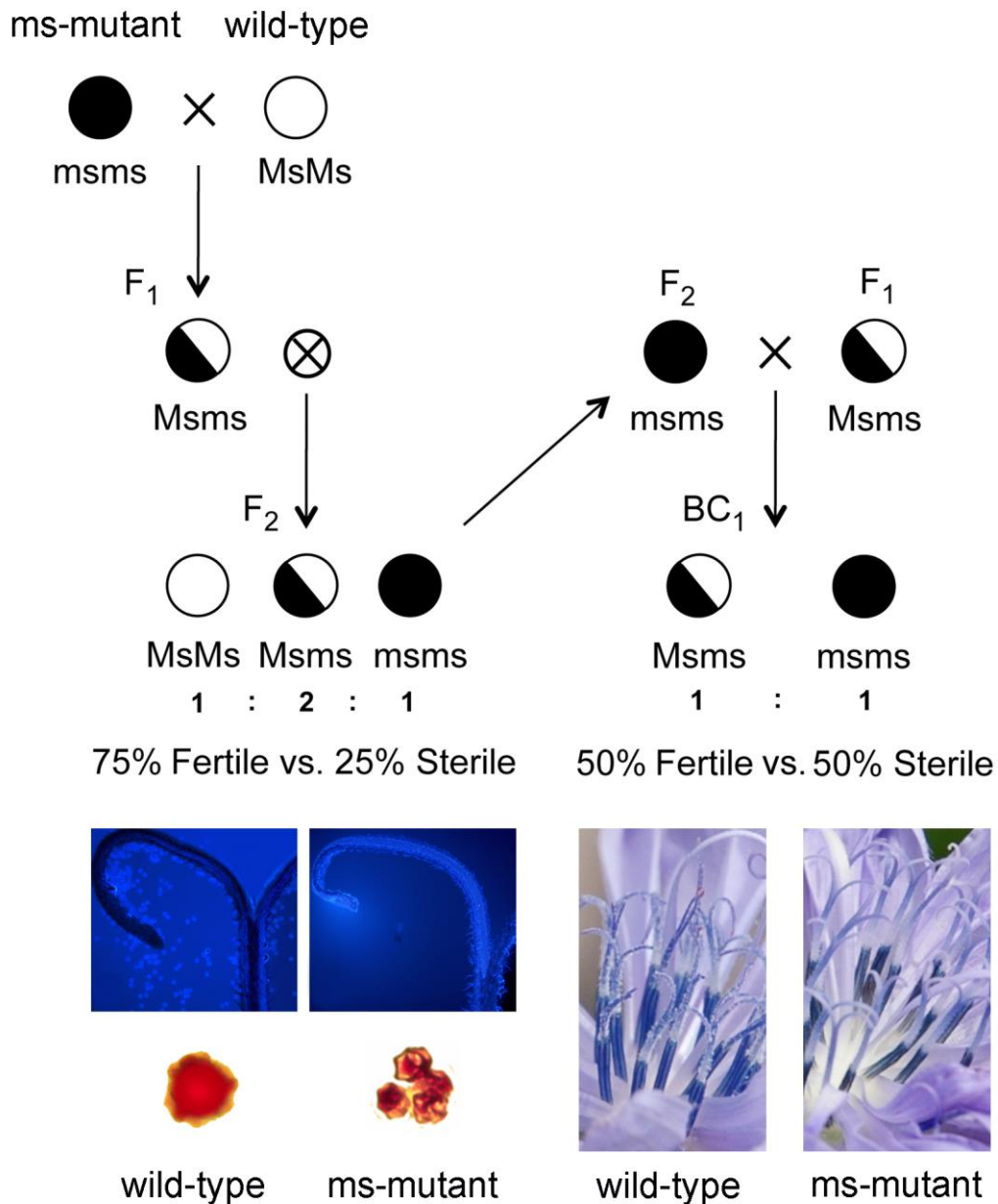


Figure 1. Schematic representation of the breeding populations developed for mapping the gene responsible for male-sterility in *Cichorium intybus*. A male-sterile mutant plant (genotype *msms*), belonging to a cultivated population of radicchio was crossed with a male-fertile plant (genotype *MsMs*) selected from a spontaneous accession of wild chicory. An F₁ male-fertile plant heterozygous at the male-sterility locus (genotype *Msms*) was selfed and an F₂ male-sterile plant (genotype *msms*) of the segregating population was then backcrossed as seed parent to a sister F₁ male-fertile plant (genotype *Msms*) used as pollen donor. Individuals of the BC₁ population were phenotyped by visual observations of anthers (*in vivo* screening for the presence/absence of pollen) and cytological investigations (*in vitro* staining of pollen using acetocarmine and DAPI solutions). At flowering, anthers were preliminary screened for the absence *vs.* presence of pollen and microscopy was then used to validate the sterile *vs.* fertile phenotype (mutants were characterized by shapeless, smaller and shrunken microspores as compared to the wild-type ones)

It is worth mentioning that in mutant plants at the stage of dehiscent anthers, microspores were arrested in their development at the uninucleate stage, and collapsed before their release from the tetrads. Viable pollen grains were never detected in mature anthers, demonstrating full expression of the male-sterility trait [15,16].

Total genomic DNA of the parents and progeny was isolated from 100 mg of fresh leaf tissue using the DNeasy[®] Plant mini-kit (Qiagen, Hilden, Germany) following the recommendations of the manufacturer. Quality and concentration of DNA samples were estimated by spectrophotometric analysis (NanoDrop 2000c UV-Vis, Thermo Fisher Scientific, San Jose, CA, USA) and quality was also assayed by agarose gel electrophoresis (1.0% w/v agarose TAE 1× gel containing 1× SYBR[®] Safe, Thermo Fisher Scientific).

Molecular mapping of the ms1 locus with simple sequence repeat (SSR) and cleaved amplified polymorphic sequence (CAPS) markers

The entire BC₁ population of leaf chicory was used for genetic mapping of the male-sterility gene using three selected SSR markers (M4.12, M4.11b and M4.10b) and one EST-derived CAPS marker previously mapped on linkage group 4 [7,15,16].

M4.12 [18] was derived from a microsatellite region (GenBank accession JF748831) in an AFLP-derived amplicon corresponding to marker E02M09 [15], whose sequence in leaf chicory was found to encompass a microsatellite region (GenBank accession JF748831), was converted into a SSR marker and renamed as M4.12 by Ghedina *et al.* [18]. This sequence-tagged site marker was mapped on linkage group 4 and was therefore included in the genetic analysis of the BC₁ population. Among the microsatellite marker loci publicly available for the leaf chicory genome [7] and associated to the linkage group 4, the M4.11b [18] (synonym EU03H01) contained an imperfect ((TG)₅CG(TG)₇) microsatellite motif (GenBank accession KF880802) and M4.10b [18] (synonym EU07G10 [7]) carried a (CT)₈TT(CT)₅CC(CT)₃TT(CT)₇ microsatellite motif (GenBank accession KX534081).

For SSR amplification, the three-primer strategy reported by Schuelke [20] was adopted, with some modifications. Briefly, forward primers were tagged at their 5' end with universal sequences (M13 or PAN2, unpublished) and used in PCR reactions in combination with sequence-specific reverse primers, and M13 and PAN2 oligonucleotides labelled with the fluorophores 6-FAM and NED, respectively.

PCR reactions were performed in a total volume of 10 μ l containing approximately 20 ng of gDNA template, 1 \times Platinum[®] Multiplex PCR Master Mix (Applied Biosystems, Carlsbad, CA, USA), GC enhancer 10% (Applied Biosystems), 0.05 μ M tailed forward primer (Invitrogen Corporation, Carlsbad, CA, USA), 0.1 μ M reverse primer (Invitrogen Corporation) and 0.23 μ M universal primer (Invitrogen Corporation). The following thermal conditions were adopted for all reactions: 2 min at 95°C for the initial denaturation step, 45 cycles at 95°C for 30 s, 55°C for 30 s and 72°C for 45 s. A final extension step at 72°C for 30 min terminated the reaction, to fill-in any protruding ends of the newly synthesized strands.

Amplicons were initially separated and visualized on 2% agarose gels in 1 \times TAE gel containing 1 \times Sybr Safe DNA stain (Life Technologies, Carlsbad, CA, USA). The remainder of the fluorescent labeled PCR products (8 μ l) was subjected to capillary electrophoresis on an ABI PRISM 3130xl Genetic Analyzer (Thermo Fisher). LIZ500 (Applied Biosystems) was used as molecular weight standard.

The CAPS marker was developed from a MADS-box gene (GenBank accession AF101420) which was initially considered [7] but later disproven [21] to be a candidate gene for male-sterility. Several primer pairs were designed for nested PCR assays using PerlPrimer v1.1.21 and used to amplify the full-length sequence and sub-regions of the MADS-box gene from plants of the BC₁ population phenotyped for male-sterility. Amplicons were mined for SNPs potentially associated with male-sterility. Amplification reactions were performed in a 9700 Thermal Cycler (Applied Biosystems) with the following conditions: initial denaturation at 94°C for 5 min followed by 30 cycles at 94°C for 30 sec, 57°C for 30 sec, 72°C for 60 sec and a final extension of 10 min at 72°C,

and then held at 4°C. The quality of PCR products was assessed on a 2% (w/v) agarose gel stained with 1× SYBR® Safe™ DNA Gel Stain (Life Technologies). Following amplification, PCR products were restricted using endonucleases (Promega, Madison, WI, USA) specific for single nucleotide variants, following the protocol suggested by the manufacturer. Amplicons were digested at 37°C for 2 hours. CAPS variants were visualized on 2.5% (w/v) agarose gels (Life Technologies) stained with 1× SYBR® Safe™ DNA Gel Stain (Life Technologies).

Segregation data from the three SSR markers and the MADS-box gene-specific CAPS marker were analyzed with JoinMap® v. 2.0 [22] using the BC₁ population type option. Genetic association between each of the markers and the male-sterility trait was assessed by recording the target *msl* locus as a qualitative trait. The grouping module was applied with a LOD threshold of 3 and a maximum recombination frequency r of 40%. The genetic distance between each pair-wise comparison of marker locus and target locus, expressed in centiMorgans (cM), was calculated from the recombination frequency corrected with the Kosambi's mapping function [23]. MapChart v.2.3 [24] was used to display the map.

Construction of a SNP-based linkage map using Genotyping-by-Sequencing (GBS)

Genomic DNA from 22 male sterile progeny and 22 male fertile progeny from the BC₁ population, along with those of the parental plants was quantified with the dsDNA BR assay on a Qubit® 1.0 fluorometer (Invitrogen, Carlsbad, CA, USA) and DNA concentrations were normalized to 20 ng/μl. DNA samples were shipped to LGC Genomics (Berlin, Germany) for GBS library preparation, sequencing and subsequent bioinformatic analysis. Briefly, DNA samples were digested with the restriction enzyme *MspI*, after which a single-stranded barcoded oligonucleotide was ligated to one side. The other side was ligated to an oligonucleotide, which was complementary to the amplification primer. Adaptor-ligated samples were separately amplified and then pooled, producing a single library. The pooled library was normalized, re-amplified and fragments between

300 bp and 500 bp were sequenced on a single lane of an Illumina NextSeq 500 v2 (2×150 bp, Illumina Inc., San Diego, CA, USA).

Raw reads were de-multiplexed and split according to their barcodes using Stacks ('process radtags' tool). After this step, reads were processed as follows: i) trimming of the 3'-end (to get a minimum average Phred quality score of 20 over a window of ten bases); ii) reads with final length < 64 bases were discarded; iii) reads with 5'-ends not matching the restriction enzyme site were discarded; iv) all reads containing undetermined (N) bases were also removed from the analysis. FastX was then used to trim ('Fastq/a trimmer' tool) the filtered data, to generate reads of the same length that, in turn, were assembled to create, by means of Stacks ('Ustacks' and 'Cstacks' tool, [25]), a set of consensus loci that represented a reference catalogue for the subsequent analysis. Bowtie2 [26] was used to align the trimmed sequences from each sample against the newly constituted reference catalogue, and GATK [27] was employed for SNP calling. The raw SNP variants were filtered by applying the following rules: i) minimum allele count exceeding eight reads; ii) allele frequency across all samples between 5% and 95%; iii) genotypes observed in at least 32 samples; iv) discard adjacent SNPs and SNPs with more than two alleles (*i.e.* only biallelic SNPs were taken into account).

Segregation data, analysed using a modified version of MapMaker v3 software [28], consisted of GBS-SNP data, genotypic scores for SSR markers M4.10b, M4.12, M2.6 and M2.4 [18] and qualitative scores for the male-sterility locus *ms1*. SSRs M2.6 and M2.4 were included because they have been shown by Gonthier *et al.* [14] to be associated with the self-incompatibility locus. Linkage groups were formed at a LOD threshold of 5. Marker orders within single linkage groups were determined using the MapMaker functions 'order', 'try' and 'ripple', and were checked manually to ensure optimal placement of the marker loci. Genetic distances were calculated using the Kosambi mapping function and graphically represented using MapChart v2.3 [24].

The newly developed genetic map was enriched by locating markers on the first genome draft of leaf chicory [19] using a default BLASTN approach (similarity>90%, E-value<1e⁻⁵⁰). Putative

functional annotation of genes present in the mapped contigs was performed using BLASTX against the TAIR10 database (E-value < 1e⁻⁵).

Validation of SNP variants linked to the male-sterility locus through Allele Specific (AS)-PCR assays

All SNPs potentially associated with the male-sterility *ms1* locus and exhibiting a maximum of three recombinant events with *ms1* were validated in a larger number of progeny through allele-specific PCR (AS-PCR) assays.

A total of 64 genomic DNA samples (*i.e.* 32 male sterile plants and 32 male sterile plants) from the same BC₁ population but not included in the GBS analysis were used for amplification using two sets of primers for each male-sterility associated SNP. Each primer set consisted of a different allele-specific forward primer and a common locus-specific reverse primer. The two allele-specific PCR primers were designed so that the 3' nucleotide was complementary to one allele of the putative polymorphism. Amplicons were separated on 1.0% w/v agarose TAE 1× gels containing 1× SYBR[®] Safe stain (Thermo Fisher Scientific). Segregation data for *ms1* and *ms1*-associated markers obtained in the entire set of 108 (44 used in GBS + 64 used in AS-PCR) BC₁ progeny were used to build a new genetic linkage map for the chromosomal block carrying the target locus with JoinMap[®] v. 2.0 [22], The BC₁ population type option was adopted and the Kosambi mapping function was used to calculate genetic distances. The resulting map was drawn with MapChart v.2.3 [24].

Results

Fine molecular mapping of the SSR and CAPS markers in the linkage group carrying the male-sterility locus

The fine genetic mapping of three selected SSR markers (Table 1) was successfully pursued and the genetic recombination estimates were validated using 198 BC₁ individual plants of the segregating population on the basis of chi-square values against independent assortment patterns. The genetic distances between the male-sterility gene and the microsatellite markers M4.12 (EU02M09) and

M4.11b (EU03H01) mapped apart from the *msI* locus were equal to 5.8 cM and 12.1 cM, respectively. An additional mapped microsatellite marker, M4.10b (EU07G10), belonging to the same linkage group, was located downstream the *msI* locus at a genetic distance of 31.2 cM (Figure 2, panels A and B).

Table 1. List of primer pairs used to amplify the mapped markers of linkage group 9 (corresponding to linkage group 4 of Cadalen *et al.* [7]). The name of the locus, the GenBank accession of the trait amplified, the polymorphism observed in ms mutants and the primer pairs are reported. SNPs marked as * were not tested with allele-specific primer combinations

Locus Name	GenBank ID	Polymorphism	Forward primer	Reverse primer
EU02M09 ³ M4.12 ²	JF748831 ³	(TC) _n	F: GGCATCGGGATAGAAAAACA	R: TCAATGCCTCAACAGAAATCC
EU03H01 ^{1,3} M4.11b ²	KF880802 ^{1,2,3}	(TG) _n CG(TG) _n	F: GCCATTCCTTCAAGAGCAG	R: AACCCAAAACCGCAACAATA
EU07G10 ¹ M4.10b ²	KX534081 ⁴	(CT) _n CATA/(CA) _n CT(CA) _n	F: CATCCATTATTGGGCAG	R: CACCAACGAACTCCTTACAAA
MADS box L2/R2 ¹ CAPS marker ⁴	AF101420 ¹ KX45584/0/1 ⁴	<i>Nco</i> I	F: TTTTGTGGGGTTTTGATTTAGA	R: TGAGATTGCATGAATGAGAACA
T11292 ⁴	KX789069 ⁴	g.205A>C	F1: GAATGAAAATTGACATAATC F2: GATGCATTAACATGGGTCT	R1: TTTGTTGATTCTGTTCCTG R2: GACTGATGGATGTCCAAT
T4403 ⁴	KX789070 ⁴	g.253A>T	F: GCGAACRAATGAGGATATATGAG	R1: GTGTTGATTGAGTGAAAATCT R2: GTGTTGATTGAGTGAAAATCA
T4390 ⁴	KX789071 ⁴	g.102C>T*; g.139T>C	F: TGTAACGCCCGTAAACCCAA	R1: GACATGTTTACTAAGGTGATGATAATATAA R2: GACATGTTTACTAAGGTGATGATAATATAG
T4391 ⁴	KX789072 ⁴	g.150C>T	F1: TAAATGTGCAATACCATGAAGC F2: TAAATGTGCAATACCATGAAGT	R: AAGTGAGTAAGTGGTTGTATTCT
T4393 ⁴	KX789073 ⁴	g.137A>G	F1: AACACATGAAGGMACTCTAG F2: AGGTCCTCATATTAAGG	R1: GATGGGTATTGAACTTATG R2: GATGTTTGTGAATGATGTTT
T4392 ⁴	KX789074 ⁴	g.68C>T	F: TTGTTGGAAGTGATGAGGTGT	R1: TGTTATTAAGTGTGTTTCGTGATATAG R2: TGTTATTAAGTGTGTTTCGTGATATAA
T4399 ⁴	KX789075 ⁴	g.24T>A	F1: AACATGATTTGTCTGCCA F2: GACATTTTTGGAACACTTTTTA	R1: CATCACAATATTCTATCCAAA R2: ACTGTTACATAATGGCTAG
T4394 ⁴	KX789076 ⁴	g.132T>C; g.235G>A*	F: GTTGGTCTGTTGATTGTTGG	R1: CATGTATATTGGAAGTTATCAACA R2: CATGTATATTGGAAGTTATCAACG
T4401 ⁴	KX789077 ⁴	g.89T>A	F1: GTAAAGAGTGGGAGTCATATA F2: GTAAAGAGTGGGAGTCATATT	R1: CTTGTTGCCTTTGTTCTAGA R2: GTTTCAACATCMTTCACT
T4402 ⁴	KX789078 ⁴	g.166T>G	F1: CAGTGGCAATGCACAT F2: GGAAGTAAATACTATACTATCGC	R1: ATAGTCTCATGGTATGCAGT R2: CATGACCACAAGGTAAACC
T4400 ⁴	KX789079 ⁴	g.231G>A	F1: CTTAATTAGGACAGAAGTAAATATAGACAG F2: CTTAATTAGGACAGAAGTAAATATAGACAA	R: GGAGTGGGGTTAAAGGGAA
T4395 ⁴	KX789080 ⁴	g.179C>T*; g.214T>C	F1: GATTCTGCAATCGGATATCGCCTTT F2: GATTCTGCAATCGGATATCGCCTTC	R: ACACCGCTGATAACCACCACC
T4398 ⁴	KX789082 ⁴	g.175C>T	F: TGCTGCCCGGAGAAATGTTAG	R1: TTTAGGCAACCGAAGTAAGGCAGTG R2: TTTAGGCAACCGAAGTAAGGCAGTA

¹[7]; ²[18]; ³[15,16]; ⁴Present paper

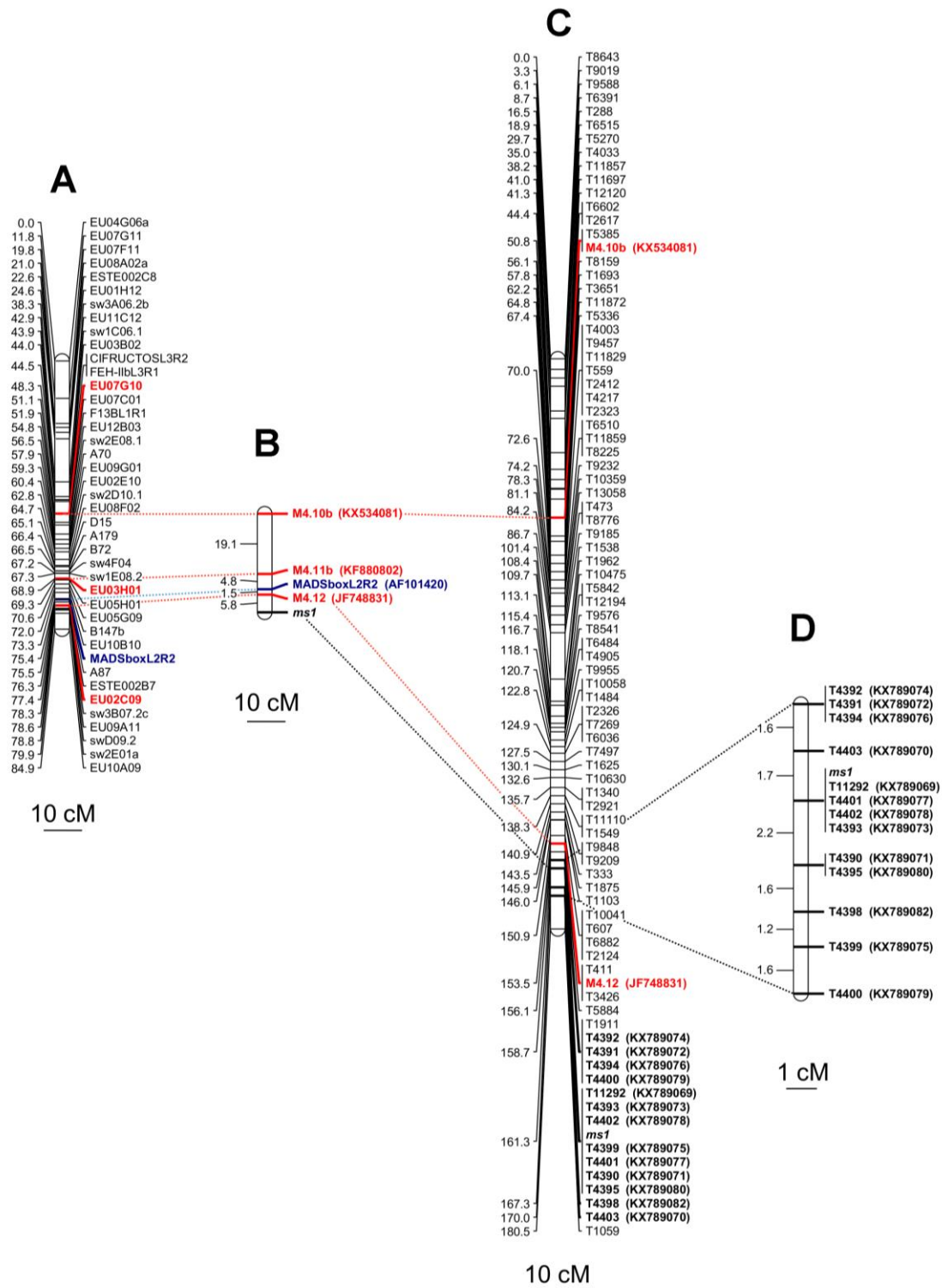


Figure 2. Molecular markers mapped on linkage group 4 and found associated to the male sterility locus (*ms1*) of *Cichorium intybus*. **A.** Linkage group 4 constructed from Cadalen *et al.* [7]. **B.** Genetic distances, expressed in cM, among three SSR markers (in red), one CAPS marker derived from the MADS-box region (in blue) and the male-sterility locus, whose estimates were calculated using a BC₁ population (198 samples) segregating 1:1 for the male-sterility trait,. The SSR markers were retrieved from Ghedina *et al.* [18] while the CAPS marker was developed in the present work. **C.** Linkage group 9 (corresponding to linkage group 4 of Cadalen *et al.* [7]) constructed using the data from the Genotyping-By-Sequencing (GBS) approach and based on 44 samples of the BC₁ population mentioned above. In red are reported the two SSR already used for the first SSR-approach, in bold are listed the 13 SNP showing ≤ 3 recombinants with the *ms1* locus. **D.** Chromosomal region around the *ms1*, considering the recombinant data of the aforesaid 13 SNP and a total number of 108 BC₁ samples

The DNA sequences of the genomic regions containing these SSR markers were deposited in GenBank as accessions KX534081, KF880802 and JF748831 (Table 2). According to the nomenclature of Cadalen *et al.* [7], the gene responsible for male-sterility was associated to linkage group 4, as predicted by Barcaccia and Tiozzo [15,16]. Sequence data published by Galla *et al.* [19] on the genome draft of leaf chicory were used to discover, predict and annotate all genes of the genomic contigs encompassing the three molecular markers used for the SSR analysis. Regarding M4.12 [18], the mapped marker closest to the *ms1* locus, its sequence was found to match with contig_84164, long 36,250 nucleotides and putatively linked with gene models AT3G11330, AT4G03600, AT5G50170 and AT5G63130 (Table 2). It is worth noting that the former one encodes for PIRL9, a member of the Plant Intracellular Ras-group-related LRRs (Leucine rich repeat proteins) and is required for differentiation of microspores into pollen grains.

Concerning the MADS-box L2/R21 gene (AF101420) as candidate, the alignment of sequences of part of its 5'-UTR region, exon 1 and the early region of intron 1 recovered from both male-sterile mutants and wild-type plants enabled to map the MADS-box locus on the linkage group 4 [7] by means a CAPS markers. In fact, three SNPs determining a restriction site were discovered in the amplified sequence of the first exon of the MADS-box gene. The cleavage site of the six-base cutter *NcoI* endonuclease was found to include a polymorphism at position 61 of the nucleotide sequence of the male fertile genotype when compared with the male sterile genotype (GenBank accessions KX455840 and KX455841). The amplification-restriction protocol for the detection of the CAPS marker alleles was applied to the total 198 BC₁ individual plants of the mapping population. In particular, 14 individuals scored recombinant genotypes when compared with the male-sterile and male-fertile phenotypes, so that the MADS-box gene was mapped at a genetic distance of 7.3 cM apart from the *ms1* locus (Figure 2, panels A and B). A genomic sequence, corresponding to contig_95308, long 8,023 nucleotides, was found to match with the sequence encompassing the CAPS marker that, from a BLASTX approach with the TAIR database, proved to be annotated as AT4G24540, encoding for a protein involved in flowering (Table 2).

Table 2. List of markers that, aligning against contigs of the first genome draft [19], were functionally annotated on the bases of matches with *Arabidopsis* protein database (TAIR10). Localization of the marker within the genic region is reported too

Marker Map Reference	GenBank ID	Marker GenBank ID	Marker localization	Gene Model Name	Predicted Function
E02M09³ M4.12²	KX639717	JF748831 ³	upstream	AT5G50170	C2 calcium/lipid-binding and GRAM domain containing protein
	KX639715		upstream	AT3G11330	Plant Intracellular Ras group-related LLR (PIRL)
	KX639716		upstream	AT5G63130	Octicosapeptide/Phox/Bem1p family-like protein
	KX639719		upstream	AT5G62230	ERECTA like protein
MADs box L2/R2¹	KX639714	AF101420 ¹	included	AT4G24540	Agamous-like 24 (AGL24)
CAPS marker⁴	KX639718	KX5455840/1 ⁴	upstream (5' UTR)	AT2G22540	SVP like protein
EU03H01^{1,3} M4.11b²	KX639711	KF880802 ^{1,2,3}	downstream	AT1G01620	Plasma membrane intrinsic like protein
EU07G10¹ M4.10b²	KX639712	KX534081 ⁴	upstream	AT5G39000	Malectin/receptor-like protein kinase family protein
T11292⁴	na	KX789069 ⁴	downstream	AT3G61700	Helicase with zinc finger domain
T4403⁴	na	KX789070 ⁴	downstream	AT4G27280	Calcium-binding EF-hand family protein
T4390⁴	na	KX789071 ⁴	na	nd	na
T4391⁴	na	KX789072 ⁴	Included (CDS – synonymous)	AT5G47910	NADPH/respiratory burst oxidase protein D (RBOHD)
T4393⁴	na	KX789073 ⁴	na	nd	na
T4392⁴	na	KX789074 ⁴	downstream	AT5G55350	MBOAT (membrane bound O-acyl transferase) family protein
T4399⁴	na	KX789075 ⁴	upstream	AT5G61890	ERF subfamily B-4, from ERF/AP2 transcription factor family.
T4394⁴	na	KX789076 ⁴	na	nd	na
T4401⁴	na	KX789077 ⁴	downstream	AT3G61420	BSD domain (BTF2-like transcription factors, Synapse-associated proteins and DOS2-like proteins)
T4402⁴	na	KX789078 ⁴	Included (CDS-non synonymous)	XP_022024718*	Transposase of the MuDR family
T4400⁴	na	KX789079 ⁴	included (intron)	AT1G15740	Leucine-rich repeat family protein
T4395⁴	na	KX789080 ⁴	upstream	AT5G56120	RNA polymerase II elongation factor
T4398⁴	na	KX789082 ⁴	Included (CDS – synonymous)	AT3G25620	ABC-2 type transporter family protein (ABCG9)

¹[7]; ²[18]; ³[15,16] ⁴Present paper; *For T4402 is reported the best match with the NR database; na=not available; nd=no significant match detected.

The first SNP-based linkage map of chicory through GBS

A GBS approach was applied to a subset of the BC₁ population, consisting of 22 male sterile and 22 male fertile plants, in order to build the first SNP-based linkage map in *Cichorium* spp. and to identify markers associated with the *ms1* locus. A raw pool of 16,353 SNPs was identified using GATK. After removal of (1) SNPs with more than 30% of missing data, (2) SNPs with a sequence depth $\leq 8\times$, (3) tri- and tetra-allelic SNPs, and (4) SNPs with allele frequencies across all samples $\leq 5\%$ and $\geq 95\%$, 1,995 SNPs were retained for the construction of a genetic linkage map. A total of 727 SNPs clustered and mapped into 9 linkage groups spanning a total length of 1,413 cM (Figure 3).

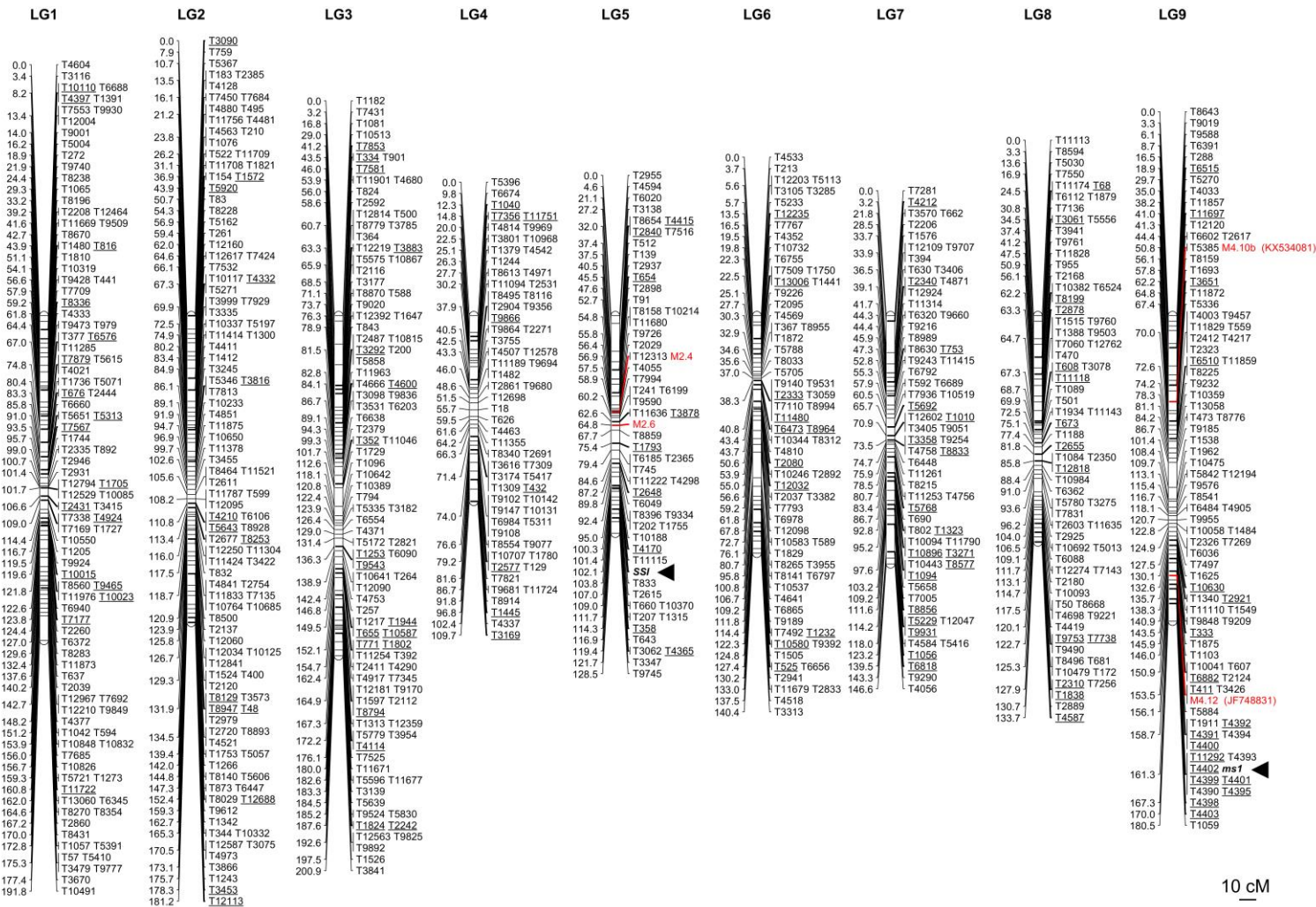


Figure 3. The first SNP-based linkage map in *Cichorium intybus*. Tags underlined represent those mapped sequences showing a significant match (BLASTN, similarity > 90%, E-value < 1e⁻⁵⁰) with contigs from the first genome draft of chicory [19] that in turn aligned against protein from the TAIR database (BLASTX, E-value < 1e⁻⁵). The correspondence between underlined tag and the best TAIR match is reported in Table 2. The male-sterility locus (*ms1*) was assessed by recording the target locus as a putative gene fully co-segregating with the trait. Four SSR markers, M4.10b, M4.12 (from linkage group 4 of Cadalen *et al.* [7]), M2.6 and M2.4 (from linkage group 2 of Cadalen *et al.* [7]) were used to genotype the same samples employed for the SNP-based map and integrated in the linkage map. M4.10b and M4.12 were chosen because co-segregating with the *ms1* [15,16], M2.6 and M2.4 were selected because were found to be associated with the sporophytic self-incompatibility (SSI) locus by Gonthier *et al.* [14]

Each mapped SNP-carrying read was used to anchor the first genome draft of leaf chicory by conducting a default BLASTN analysis. A total of 688 out of 727 genetically mapped reads (95%) strongly matched (similarity>90%, E-value<1e⁻⁵⁰) sequences in at least one genomic contig. Considering top hits only, some 3.7 Mb (0.3% of the whole genome) of the chicory genome sequence was anchored. Among the mapped contigs, 18.6% aligned with expressed regions from the TAIR Database (BlastX, E-value<1e⁻⁵). This allowed organizing 128 coding regions over the 9 linkage groups (Table 3). Among them, it was possible to identify 46 different enzymatic proteins, 8 membrane proteins and 4 transcription factors.

The *msl* locus, along with the M4.10b and M4.12 SSR markers, mapped to linkage group 9 of our genetic map (Figure 3 and Figure 2, panel C), allowing us to associate it with the linkage group 4 from Cadalen *et al.* [7] (see also Figure 2, panel A). In particular, M4.12 co-segregated with the target gene and it was mapped at 7.8 cM from the *msl* locus (Figure 2, panel C). As many as Thirteen SNPs exhibited ≤ 3 recombination events with the target *msl* locus, seven of which (T11292, T4393, T4402, T4399, T4401, T4390 and T4395) cosegregated with *msl* in the population of 44 BC₁ progeny (Figure 2, panel C). The GBS reads corresponding to these 13 markers have been deposited in GenBank under accession numbers KX789069-KX789080, KX789082. Ten of the chicory contigs carrying the mapped SNPs showed a significant match (E-value<1e⁻⁵) with TAIR database (Table 2).

Since a recent study located the SSI locus [14] in linkage group 2 from Cadalen *et al.* [7], two SSR markers from this group (namely M2.6 and M2.4) were used to genotype the BC₁ samples employed for the GBS strategy. This analysis allowed us to associate the S-locus of leaf chicory to our linkage group 5 and the genetic distance between the two SSR markers resulted equal to 7.9 cM (Figure 3).

Table 3. List of GBS-derived tags mapped over the 9 linkage groups (LG) of *Cichorium intybus* subsp. *intybus* var. *foliosum* and hypothetic function on the basis of matches with *Arabidopsis* protein database (TAIR database). Tags underlined were found to be associated at 0cM with the male sterility locus (*ms1*) based on 108 samples. For T4402 (LG9) is reported the best match with the NR database

LG1		LG2		LG3		LG4		LG5		LG6		LG7		LG8		LG9	
Tag	TAIR	Tag	TAIR	Tag	TAIR	Tag	TAIR	Tag	TAIR	Tag	TAIR	Tag	TAIR	Tag	TAIR	Tag	TAIR
T10110	AT4G23160	T3090	AT5G27870	T7853	AT4G22270	T1040	AT1G63800	T4415	ATMG00300	T12235	AT5G40170	T4212	AT3G04650	T68	AT2G27920	T6515	AT4G32090
T4397	AT1G59830	T1572	AT3G47680	T334	ATMG00820	T7356	AT1G20960	T2840	AT4G23160	T13006	ATMG00300	T2340	AT2G28550	T3061	AT5G22360	T11697	AT4G23160
T816	ATMG00300	T5920	ATCG00780	T7581	AT2G17220	T11751	AT1G12480	T654	AT4G23160	T2333	ATMG00750	T753	AT4G23160	T8199	AT3G44160	T3651	AT1G21280
T8336	AT1G41920	T4332	AT1G31180	T3883	AT3G25800	T9866	AT5G37020	T3878	AT4G23160	T11480	ATMG00750	T5692	AT5G20870	T2878	AT4G23160	T6510	AT5G48050
T6576	AT5G63540	T3816	AT4G17250	T3292	AT3G03960	T432	AT4G23160	T1793	AT3G49142	T6473	ATMG00810	T1010	AT5G48050	T608	AT5G48050	T10630	AT5G48100
T7879	AT4G23160	T4210	AT4G23160	T4600	AT4G35160	T2577	AT2G46495	T2648	AT5G18200	T8964	ATMG00300	T3358	AT4G03230	T11118	AT5G56490	T2921	AT1G58200
T676	AT3G21250	T5643	ATCG00740	T352	ATMG00300	T1445	AT5G35160	T4170	AT1G80350	T2080	ATMG00300	T8833	AT5G57140	T673	AT4G23160	T333	AT1G79000
T5313	AT1G16800	T8253	ATMG00300	T1253	ATMG00810	T3169	ATMG00750	T358	AT1G74680	T12032	ATMG00300	T5768	AT3G02645	T2655	AT1G34070	T6882	AT5G42810
T7567	AT4G29840	T8129	AT2G19880	T9543	ATMG00300			T4365	AT3G29638	T1232	AT4G23160	T1323	AT1G31540	T12818	AT2G05710	T411	AT5G25900
T1705	AT3G29785	T8947	AT5G19690	T1944	AT3G49142					T10580	AT4G33070	T10896	AT5G66910	T9753	AT4G15040	T4392	AT5G55340
T2431	AT4G23160	T48	AT4G23160	T655	ATMG01250					T525	AT2G41770	T3271	AT1G17210	T7738	AT2G33680	T4391	AT5G47910
T4924	AT4G23160	T12688	AT5G41980	T10587	AT4G23160							T8577	AT3G14590	T2310	AT5G52520	T4400	AT1G15740
T10015	ATMG00300	T3453	AT4G03080	T771	AT5G48050							T1094	AT5G11150	T1838	AT5G05200	<u>T11292</u>	AT3G61700
T9465	AT1G21580	T12113	ATMG00300	T1802	AT3G47210							T8856	ATMG00300	T4587	AT2G40280	<u>T4402</u>	XP_022024718
T10023	AT5G41980			T8794	AT4G30790							T5229	AT2G15220			T4399	AT5G61890
T7177	AT4G23160			T4114	AT5G20320							T9931	ATMG00300			<u>T4401</u>	AT3G61420
T11722	ATMG00300			T2242	AT3G04350							T1056	AT5G49630			T4395	AT5G56120
				T1824	ATMG00820							T6818	AT4G23160			T4398	AT3G25620
																T4403	AT4G27280

Validation of the SNPs linked to the *ms1* locus

The map positions of the 13 male sterility-associated SNPs were validated by analyzing an additional 64 BC₁ progeny (32 male sterile and 32 male fertile). This brought the number of BC₁ progeny analyzed to 108, including the initial pool of 44 BC₁ samples analyzed by GBS. For each of the SNP markers, two pairs of primers targeting the two alleles were used in separate reactions. Because BC₁ progeny are either heterozygous or homozygous for the recurrent parent allele, the primer set that amplified the recurrent parent allele generated amplification products in all BC₁ progeny and hence acted as positive control. The other primer set amplified the alternate allele which was present only in heterozygous progeny (Figure 4).

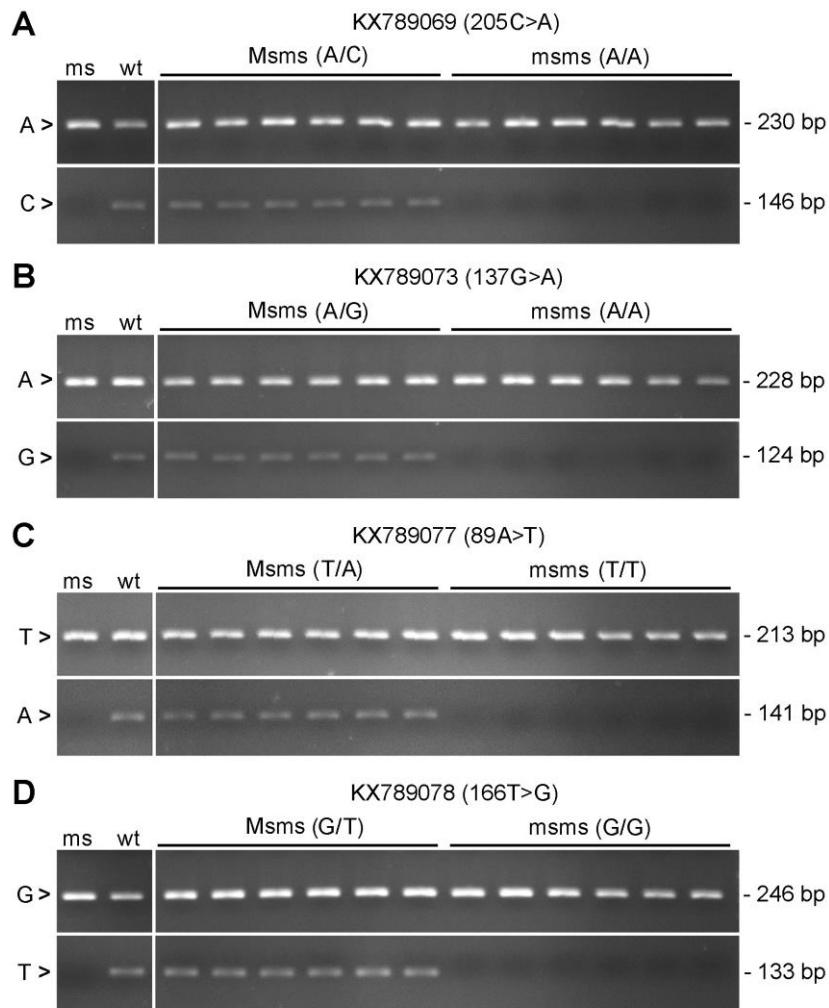


Figure 4. A summary of the AS-PCR profiles generated in the male sterile mutant (*msms*) and male fertile wild-type (*Msms*) parents and a representative subset of 12 BC₁ progeny plants (segregating 1 *Msms* : 1 *msms*) with the allele-specific primers for the four diagnostic SNP markers co-segregating with male-sterility. For each SNP marker is indicated the point mutation that discriminates male-sterile and male-fertile plants, the specific length of amplicons and the GenBank accession number.

Linkage analysis across the 108 progeny showed that all 13 SNPs mapped to a 9.9 cM region on linkage group 9 (corresponding to linkage group 4 according to Cadalen *et al.* [7], Figure 2, panel D). Four SNPs (T11292, T4393, T4401 and T4402) co-segregated with the *ms1* locus (Figure 2, panel D).

The GBS reads carrying the SNP markers T11292 (KX789069) and T4401 (KX789077) aligned against contig_55191 and contig_71514 in the chicory draft genome assembly [19]. These contigs had significant matches with gene models AT3G61700 (contig_55191) and AT3G61420 (contig_71514) (Table 2 and Table 3) which encoded, respectively, for a helicase with zinc-finger protein and a BSD domain (BTF2-like transcription factors, synapse-associated proteins and DOS2-like proteins). The SNPs were located downstream from the coding region. The read carrying SNP T4402 (KX789078) aligned with contig_2496, but this sequence did not show any significant blast hits with proteins in the TAIR database. Extending the BLASTX analyses to NCBI's NR database, contig_2496 was found to match significantly (E-value= $1e^{-43}$, Similarity 44%) with a locus from *Helianthus annuus* (Asteraceae family) encoding for a transposase of the MuDR family (XP_022024718). Moreover, the discriminant SNP resulted to be non-synonymous and it was located in the zinc-finger domain (ZNF_PMZ) of the protein. The read carrying SNP T4393 (KX789073) did not show any significant match with either TAIR or the NR database.

Discussion

To the best of our knowledge, this is the first time that male-sterile mutants have been genetically characterized in leaf chicory (*Cichorium intybus* subsp. *intybus* var. *foliosum*). The main goals of our study were to build the first SNP-based linkage map in this species and to accomplish the molecular mapping of the male-sterility trait.

A deep comprehension of nuclear male-sterility systems is extremely important for their exploitation in plant breeding, as male-sterility is one of the most effective methods to produce F₁ hybrid varieties in crop plants [29–32]. F₁ hybrids are usually developed by crossing two highly

homozygous parental lines, selected to obtain highly heterozygous progeny, which are characterized not only by high uniformity of phenotypic traits, but also by strong heterosis in terms of productivity.

A recent cytological study of a naturally occurring male-sterile mutant in leaf chicory has shown that micro-sporogenesis proceeds regularly up to the development of tetrads [15,16]. After that, all microspores arrest their developmental program. At the beginning of micro-gametogenesis, non-viable shrunken microspores were clearly visible within anthers. Moreover, detailed investigations indicated the occurrence of meiotic abnormalities in the male-sterile mutants, especially at prophase I. Abnormal pairings and chromosomal loops were observed during pachytene. This new mutant, whose male-sterility is caused by a recessive nuclear gene, has been applied in the production of F₁ hybrids of Radicchio, and has been recently subjected to patenting [15,16]. However, beyond the fact that the NMS was discovered and mapped in root chicory [11,14] and leaf chicory [15,16] within linkage groups 5 and 4, respectively, no genetic information is available about this locus.

Exploiting an SSR-based approach, the gene responsible for male-sterility in leaf chicory was found genetically linked to the genomic locus M4.12 (JF748831), an AFLP-derived amplicon encompassing an SSR region, about 5.8 cM apart from the *msI* locus. Two additional SSR markers, corresponding to genomic loci M4.10b (KX534081) and M4.11b (KF880802), were mapped in the same linkage group at a genetic distance equal to, respectively, 31.2 cM and 12.1 cM from the *msI* locus. A marker-assisted selection (MAS) approach can be fully exploited in plant breeding programs if selectable SSR markers are tightly linked to the target locus, possibly mapped on both upstream and downstream of the gene of interest. Unfortunately, the three discovered SSR markers were found loosely linked to the target locus and all of them were genetically mapped downstream the *msI* gene, hence barely useful for MAS applications. Considering the functional annotation of the three SSR-containing genomic sequences, E02M09, the closest marker to the *msI* locus, was found to be putatively linked with a TAIR gene (AT3G11330, PIRL 9 protein) required for differentiation of microspores into pollen grains [33], which definitely is the phenomenon described

as disrupted in our *msl* mutants based on cytological observations [15,16]. However, the relatively high number of recombinants detected in the BC₁ population, supporting a genetic distance of 5.8 cM from the *msl* locus, proved that it could not be responsible for male-sterility in leaf chicory.

The diagnostic CAPS marker derived from the MADS-box L2/R2 gene (KX455840-KX455841) was found genetically linked at 7.3 cM apart from the *msl* locus. This marker encompasses a genomic block that includes a MADS-box protein (AT4G24540) and a protein that acts as a floral repressor (AT4G22540). Functional analyses by molecular genetic studies in model eudicots, such as *Arabidopsis thaliana* L., have shown that the proteins encoded by these two genes are both essential for the regulation of various aspects of flower development [34] but no information about their involvement in the male-sterility mechanism is available. Moreover, the relatively high number of recombinants detected in the BC₁ population raises some doubts regarding its role in the pollen development in leaf chicory.

A genotyping-by-sequencing approach allowed us to construct the first SNP-based linkage map and to narrow down the genomic window around the *msl* locus in leaf chicory. A total of 727 reads-carrying SNPs were clustered into 9 linkage groups. The first genome draft of leaf chicory was then used to anchor 688 contigs and 128 coding regions to the genetic map. Among the enzymes mapped, it was possible to identify proteins involved in the biosynthesis of N-glycan (AT5G19690), cutin, suberin and wax (AT5G55340), diterpenoid (AT5G25900) and amino acids, including valine, leucine and isoleucine (AT1G31180). Other enzymes resulted specifically involved in metabolism processes like glycine, serine and threonine metabolism (AT4G29840), glycoxylate and dicarboxylate metabolism (AT2G05710), inositol phosphate (AT5G42810) and ascorbate and aldarate metabolism (AT5G56490). Two different proteins VAMP713 (AT5G11150) and VAMP714 (AT5G22360), with a key role in the vacuolar trafficking during salt stress [35] were mapped in the linkage groups 7 and 8, respectively. Finally, a noteworthy SNP marker mapped in the linkage group 9 was associated to a genomic contig that, in turn, matched with an ERF BUD ENHANCER in *Arabidopsis* (EBE, AT5G61890). This transcription factor, member of the

APETALA2/ETHYLENE RESPONSE FACTOR (AP2/ERF) transcription factor superfamily, was found: i) to promote cell proliferation, leading to enhanced callus growth; ii) to stimulate axillary bud formation and outgrowth; iii) to affect shoot branching, acting in cell cycle regulation and dormancy breaking [36].

Curiously, 22 mapped contigs matched with as many mitochondrial genes of *Arabidopsis* and, in particular, ATMG00300 (14 matches with as many contigs) and ATMG00750 (3 matches with as many contigs). From TAIR database they resulted annotated as ‘Gag-Pol-related retrotransposon family protein’ and ‘Gag-Pol-Env polyprotein’, respectively. We found that these two classes of mitochondrial retrotransposons were detected in multiple copies throughout the nuclear genome of several species. At this regards, according to what reported in GenBank database, at least three ATMG00300-like copies were located within chromosomes 3, 7 and 15 of *Malus domestica* as well as in the linkage groups 2, 5 and 6 of *Glycine max*. In *Arabidopsis thaliana*, the same locus was found also within chromosome 2. This is in accordance with large and unexpected organellar-to-nuclear gene-transfer events highlighted in species like rice [37] and *Arabidopsis* [38]. Moreover, the abundance of retrotransposon sequences identified within the chicory map (*e.g.* 24 out of 128 coding regions were retrotransposon family proteins) is coherent with what already reported in other species like rice or maize where the retroelements represented, respectively, 14% [39] and 49% [40] of the whole genome.

Two SSR markers (M2.6 and M2.4), mapped on the linkage group 2 by Cadalen *et al.* [7], known for carrying the SSI locus [14] in root chicory, were used to genotype the 44 samples employed for the GBS. This analysis enabled to associate the self-incompatibility locus to the linkage group 5 of leaf chicory. The genetic distance between these two mapped SSR markers was 7.9 cM whereas they were 14.9 cM apart in the genetic map developed by Cadalen *et al.* [7].

According to Chen [41], plant receptor-like protein kinases (RLKs) are classified according to sequence motifs in the putative extracellular receptor domains. One of the most represented RLKs families is the SRK group (S-locus Receptor Kinase) characterized by an S-domain rich in cysteine

residues. In the Brassicaceae taxon, self-incompatibility is controlled by the SRK gene, along with the S-locus glycoprotein (SLG) gene as enhancer, expressed in the stigma (female determinant) and by the male determinant (SP11/SCR) expressed in the anther. These genes are tightly linked at the S-locus [42]. For this reason, a search for SRKs was performed among the mapped coding regions. A total of 18 coding regions mapped all over the 9 linkage groups proved to match with the AT4G23160 locus (cysteine-rich receptor-like protein kinase 8). This finding is not surprising given that more than 300 different RLKs were mapped in Arabidopsis. In particular, three cysteine-rich receptor-like protein kinases co-segregated with marker loci clustered in our linkage group 5 (linkage group 2 according to Cadalen *et al.* [7]), the one carrying the self-incompatibility locus. According to Gonthier *et al.* [14], the SSI locus is loosely linked to the SSR marker coded as M2.4 being positioned at 45.1 cM apart, whereas two candidate regions, namely T654 and T3878, could be located most likely at 11.4 cM (downstream) and 7.3 cM (upstream), respectively, from the target locus. Further genetic analyses are required to confirm whether one of these loci could be selected as candidate gene for the SSI system.

Recording the target *msl* locus as a putative gene fully co-segregating with the trait mapped and using two SSR markers (*i.e.* M4.10b and M4.12, according to Ghedina *et al.* [18]) co-segregating with the *msl*, enabled to overlap our linkage group 9 with linkage group 4 from Cadalen *et al.* [7] (see Figure 2, panels A-C). At least 13 SNPs were found tightly linked to the target locus, exhibiting three or less recombinants (Figure 2, panel C). In particular, seven out of 13 SNPs (*i.e.* T11292, T4393, T4402, T4399, T4401, T4390 and T4395) did not show recombinants. To increase the robustness of this finding, an AS-PCR assay was developed focusing on these 13 loci and increasing the number of BC₁ samples assayed up to 108. This new round of analysis proved to be fast, cheap and highly efficient and, most importantly, allowed us to finely map all the SNPs in a chromosomal DNA region spanning 9.9 cM (Figure 2, panel D) in length, both upstream and downstream the *msl* locus. The allelic variants of four SNPs proved still to fully co-segregate with the target trait and the corresponding reads were mapped at 0 cM from the *msl*. Interestingly, two

of these genomic sequences, namely T11292 (KX789069) and T4401 (KX789077), retrieved a significant match with two TAIR gene models: a helicase with zinc-finger protein (AT3G61700) and a BSD domain (BTF2-like transcription factors, synapse-associated proteins and DOS2-like proteins, AT3G61420) characterizing the RNA polymerase II transcription factor B. The fact that these two genes resulted to be strictly associated also in the chromosome 3 of *Arabidopsis thaliana* (~100 kb one from the other), strengthens the possibility that they may interact synergistically. Moreover, according to Honys and Twell [43], both genes resulted differentially expressed during micro-gametogenesis. Nevertheless, no information about their involvement in the male-sterility mechanism is available and the position of the two SNPs downstream of the coding regions raises some doubts regarding their role in the pollen development.

Of much more interest was T4402 (KX789078). It was found to align against contig_2496 that, in turn, matched with a locus encoding for a transposase of the MuDR family (XP_022024718) from *Helianthus annuus* (Asteraceae family). The SNP resulted to be non-synonymous and located in the zinc-finger domain (ZNF-PMZ) of the protein. According to Raizada *et al.* [44], transcripts from MuDRB and MuDRA genes increase significantly in mature pollen. Additionally, they demonstrated that the promoters of these two genes, located in transposon terminal inverted repeats (TIRs) of the MuDR, contain functionally defined pollen enhancers. According to this, is it feasible that the non-synonymous SNP differentially detected within the male sterile and male fertile accessions and specifically located in the zinc finger domain (ZNF-PMZ) could affect the structure and thus the functionality of the MuDR protein. It is well known that the central function of synapsis is the recognition of homologues by pairing, an essential step for a successful meiosis and that irregular synapsis for some of the homologous chromosomes may alter the further development of microspores, leading to the failure of gametogenesis. At this regard, a study in maize [45] proved that MuDR transposase increases the frequency of meiotic crossovers in the vicinity of a Mutator transposable element (Mu) insertion. We can hypothesize that one or more mutations within the previously mentioned protein may produce those abnormal pairings and chromosomal loops that

were observed during pachytene (when chromosomal crossing-over occurs) in the male sterile mutants. In details, all male-sterile mutants revealed different types of meiotic defects, including homolog miss-pairing at prophase I, along with chromatin bridges observed in ana-telophase II. Homologues were not completely pairing each other at pachytene stage and aberrant structures characterized by several loops, due to partial or aspecific pairing between chromosomes, were often observed. Moreover, several cases of chromatin bridges, *i.e.* bridges made of chromatin occurring between newly forming cells, were found in the male-sterile mutants [15,16].

Approaching the MuDR from another point of view and considering its transposition activity, we cannot exclude that male-sterility is the result of one or more Mu insertion events within a functional gene involved in pollen development. The first male sterile mutant produced in *Arabidopsis thaliana* represents a fascinating example, having being obtained through the insertion of a transposable element in the 3'-end of a gene later named MALE STERILITY 2 [46]. Although further studies will be necessary to confirm whether the SNPs mapped at 0 cM from the *ms1* locus maintain the same segregation pattern extending the analysis to other leaf chicory genetic backgrounds, these easily detectable DNA markers are immediately exploitable to implement rapid diagnostic essays that will find utility for marker-assisted selection programs.

In conclusion, the male-sterility gene (*ms1*) of leaf chicory was firstly confined to a chromosomal region spanning 5.8 cM through an SSR-based approach. The construction of a SNP-based linkage map and the application of an AS-PCR assay enabled to narrow down the genomic window around the target locus and to select SNVs mapped at 0 cM from *ms1* locus. Moreover, the newly developed genetic linkage map allowed us select a transposase of the MuDR family fully co-segregating with the *ms1* locus, which could be considered a primary candidate gene. Overall, our findings will be crucial to identify, clone and characterize the gene responsible for male sterility in leaf chicory.

References

1. Kaur S, Cogan NOI, Stephens A, Noy D, Butsch M, Forster JW, et al. EST-SNP discovery and dense genetic mapping in lentil (*Lens culinaris* Medik.) enable candidate gene selection for boron tolerance. *Theor Appl Genet.* 2014;127(3):703–13.
2. Zhao X, Han Y, Li Y, Liu D, Sun M, Zhao Y, et al. Loci and candidate gene identification for resistance to *Sclerotinia sclerotiorum* in soybean (*Glycine max* L. Merr.) via association and linkage maps. *Plant J.* 2015;82(2):245–55.
3. Huang S, Liu Z, Yao R, Li D, Zhang T, Li X, et al. Candidate gene prediction for a petal degeneration mutant, pdm, of the Chinese cabbage (*Brassica campestris* ssp. *pekinensis*) by using fine mapping and transcriptome analysis. *Mol Breed.* 2016;36(3):1–10.
4. Colasuonno P, Gadaleta A, Giancaspro A, Nigro D, Giove S, Incerti O, et al. Development of a high-density SNP-based linkage map and detection of yellow pigment content QTLs in durum wheat. *Mol Breed.* 2014;34(4):1563–78.
5. Schumann MJ, Zeng Z-B, Clough ME, Yencho GC. Linkage map construction and QTL analysis for internal heat necrosis in autotetraploid potato. *Theor Appl Genet.* 2017;130(10):2045–56.
6. Marcotuli I, Gadaleta A, Mangini G, Signorile AM, Zacheo SA, Blanco A, et al. Development of a high-density SNP-based linkage map and detection of QTL for β -Glucans, Protein Content, Grain yield per spike and heading time in durum wheat. *Int J Mol Sci.* 2017;18(6).
7. Cadalen T, Mörchen M, Blassiau C, Clabaut A, Scheer I, Hilbert JL, et al. Development of SSR markers and construction of a consensus genetic map for chicory (*Cichorium intybus* L.). *Mol Breed.* 2010;25(4):699–722.
8. Barcaccia G, Varotto S, Soattin M, Lucchin M, Parrini P. Genetic and molecular studies of sporophytic self-incompatibility in *Cichorium intybus*. In: *L Proc of the Eucarpia meeting on Leafy Vegetables Genetics and Breeding.* Noordwijkerhout, The Netherland; 2003. p. 154.
9. Lucchin M, Varotto S, Barcaccia G, Parrini P. Chicory and Endive. In: Prohens-Tomás J, Nuez F, editors. *Handbook of Plant Breeding, Vegetables I: Asteraceae, Brassicaceae, Chenopodiaceae.* New York, USA; 2008. p. 1–46.
10. Pécaut P. Etude sur le système de reproduction de l'endive (*Cichorium intybus* L.). *Ann Amélior Plantes.* 1962;12:265–96.
11. Desprez B, Delesalle L, Dhellemmes C, Desprez. Genetics and breeding of industrial chicory. *Comptes Rendus l'Academie d'Agriculture Fr.* 1994;80:47–62.
12. Desprez F, Bannerot H. A study of pollen tube growth in witloof chicory. In: *Proc of the Eucarpia meeting on leafy vegetables.* Littlehampton, UK; 1980. p. 47–52.
13. Eenink AH. Compatibility and incompatibility in witloof-chicory (*Cichorium intybus* L.). 3. Gametic competition after mixed pollinations and double pollinations. *Euphytica.* 1982;31(3):773–86.
14. Gonthier L, Blassiau C, Mörchen M, Cadalen T, Poiret M, Hendriks T, et al. High-density genetic maps for loci involved in nuclear male sterility (NMS1) and sporophytic self-incompatibility (S-locus) in chicory (*Cichorium intybus* L., Asteraceae). *Theor Appl Genet.* 2013;126(8):2103–21.
15. Barcaccia G, Tiozzo Caenazzo S. New male sterile *Chicorium* spp. mutant, parts or derivatives, where male sterility is due to a nuclear recessive mutation linked to a polymorphic genetic marker, useful for producing mutant F1 hybrids of *Chicorium* spp. WO2012163389-A1, 2012.
16. Barcaccia G, Tiozzo Caenazzo S. New male sterile mutant of leaf chicory, including Radicchio, used to produce chicory plant and seeds with traits such as male sterility exhibiting cytological phenotype with shapless, small and shrunken microspores in dehiscent anthers. US2014157448-A1, 2014.
17. Muys C, Thienpont CN, Dauchot N, Maudoux O, Draye X, Cutsem P Van. Integration of AFLPs, SSRs and SNPs markers into a new genetic map of industrial chicory (*Cichorium intybus* L. var.

- sativum). *Plant Breed.* 2014;133(1):130–7.
18. Ghedina A, Galla G, Cadalen T, Hilbert J-L, Caenazzo ST, Barcaccia G. A method for genotyping elite breeding stocks of leaf chicory (*Cichorium intybus* L.) by assaying mapped microsatellite marker loci. *BMC Res Notes.* 2015;8(1):831.
 19. Galla G, Ghedina A, Tiozzo SC, Barcaccia G. Toward a First High-quality Genome Draft for Marker-assisted Breeding in Leaf Chicory, Radicchio (*Cichorium intybus* L.). In: Abdurakhmonov IY, editor. *Plant Genomics.* Rijeka: InTech; 2016.
 20. Schuelke M. An economic method for the fluorescent labeling of PCR fragments A poor man ' s approach to genotyping for research and high-throughput diagnostics . *Nat Biotechnol.* 2000;18:233–4.
 21. Barcaccia G, Ghedina A, Lucchin M. Current Advances in Genomics and Breeding of Leaf Chicory (*Cichorium intybus* L.). *Agriculture.* 2016;6(4):50.
 22. Stam P. JoinMap 2.0 deals with all types of plant mapping populations. In: *Plant and animal genomics III Conference.* San Diego, California; 1995.
 23. Kosambi D. The estimation of map distances from recombination values. *Ann Eugen.* 1943;12(1):172–5.
 24. Voorrips R. MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. *J Hered.* 2002;93(1):77–8.
 25. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: An analysis tool set for population genomics. *Mol Ecol.* 2013;22(11):3124–40.
 26. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie2. *Nat Methods.* 2013;9(4):357–9.
 27. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–303.
 28. Lincoln SE, Daly MJ, Lander ES. Constructing Genetic Linkage Maps with MAPMAKER/EXP Version 3.0: A Tutorial and Reference Manual. Whitehead Institute for Biomedical Research. 1993. p. 1–49.
 29. Acquaah G. Breeding hybrid cultivars. In: *Principles of plant genetics and breeding.* 2nd ed. Chichester, UK; 2012. p. 355–73.
 30. Barclay A. Hybridizing the world. *Rice Today.* 2010;9:32–35.
 31. Rajeshwari R, Sivaramakrishnan S, Smith RL, Subrahmanyam NC. RFLP analysis of mitochondrial DNA from cytoplasmic male sterile lines of pearl millet. *Theor Appl Genet.* 1994;88:441–8.
 32. Havey M. The use of cytoplasmic male sterility for hybrid seed production. In: Daniell H, Chase C, editors. *Molecular Biology and Biotechnology of Plant Organelles.* Berlin, Heidelberg, New York: Springer; 2004. p. 623–34.
 33. Forsthoefel NR, Klag KA, Simeles BP, Reiter R, Brougham L, Vernon DM. The Arabidopsis Plant Intracellular Ras-group LRR (PIRL) Family and the Value of Reverse Genetic Analysis for Identifying Genes that Function in Gametophyte Development. *Plants (Basel).* 2013;2(3):507–20.
 34. Yamaguchi T, Hirano H-Y. Function and Diversification of MADS-Box Genes in Rice. *Sci World J.* 2006;6:1923–32.
 35. Leshem Y, Melamed-Book N, Cagnac O, Ronen G, Nishri Y, Solomon M, et al. Suppression of Arabidopsis vesicle-SNARE expression inhibited fusion of H₂O₂-containing vesicles with tonoplast and increased salt tolerance. *Proc Natl Acad Sci U S A.* 2006;103:18008–13.
 36. Mehrnia M, Balazadeh S, Zanol M-I, Mueller-Roeber B. EBE, an AP2/ERF transcription factor highly expressed in proliferating cells, affects shoot architecture in Arabidopsis. *Plant Physiol.* 2013;162(2):842–57.

37. Ueda M, Tsutsumi N, Kadowaki K ichi. Translocation of a 190-kb mitochondrial fragment into rice chromosome 12 followed by the integration of four retrotransposons. *Int J Biol Sci.* 2005;1(3):110–3.
38. Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, et al. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature.* 1999;402(6763):761–8.
39. Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, et al. The genome sequence and structure of rice chromosome 1. *Nature.* 2002;420(November):312–6.
40. Meyers BC, Tingey S V, Morgante M, Meyers BC, Tingey S V, Morgante M. Abundance , Distribution , and Transcriptional Activity of Repetitive Elements in the Maize Genome. *Genome Res.* 2001;11:1660–76.
41. Chen Z. A Superfamily of Proteins with Novel Cysteine-Rich Repeats. *Plant Physiol.* 2001;126(2):473–6.
42. Iwano M, Takayama S. Self/non-self discrimination in angiosperm self-incompatibility. *Curr Opin Plant Biol.* 2012;15(1):78–83.
43. Honys D, Twell D. Transcriptome analysis of haploid male gametophyte development in *Arabidopsis*. *Genome Biol.* 2004;5(11):R85.1-R.85.13.
44. Raizada MN, Benito MI, Walbot V. The MuDR transposon terminal inverted repeat contains a complex plant promoter directing distinct somatic and germinal programs. *Plant J.* 2001;25(1):79–91.
45. Yandea-Nelson MD, Zhou Q, Yao H, Xu X, Nikolau BJ, Schnable PS. MuDR transposase increases the frequency of meiotic crossovers in the vicinity of a Mu insertion in the maize *a1* gene. *Genetics.* 2005;169(2):917–29.
46. Aarts MGM, Dirkse WG, Stiekema WJ, Pereira A. Transposon tagging of a male sterility gene in *Arabidopsis*. *Nature.* 1993;363(6431):715–7.